Effects of Mental Model Quality on Collaborative System Performance

A Thesis
Presented to
The Academic Faculty

by

Bart D Wilkison

In Partial Fulfillment
of the Requirements for the Degree
Master of Science in the
School of Psychology

Georgia Institute of Technology
May 2008

EFFECTS OF MENTAL MODEL QUALITY ON COLLABORATIVE SYSTEM PERFORMANCE

Approved by:

Dr. Arthur D. Fisk, Advisor
School of Psychology
Georgia Institute of Technology

Dr. Wendy A. Rogers
School of Psychology
Georgia Institute of Technology

Dr. Gregory M. Corso
School of Psychology
Georgia Institute of Technology

Date Approved: 11 March 2008

ACKNOWLEDGEMENTS

First, I would like to thank my wife, Brooke and son, Brady for their support through these two years of graduate school. Thanks to my parents for their continued support. I want to also thank the US Army and the department of Behavioral Sciences and Leadership at the Unites States Military Academy for allowing me this tremendous opportunity. The members of the Human Factors and Aging Laboratory are also deserving of thanks, without your guidance, help, and support this would not have been possible. Thanks to all the Georgia Tech students who volunteered for this study and provided the data. G. Stepanhov was kind enough to provide the programming expertise to turn my ideas into a functional program. Thanks to Dr. Roberts and Hi Shin Shim for statistical guidance. Finally, I want to acknowledge the efforts of Dr. Fisk and Dr. Rogers; you allowed me to concentrate on this research and the completion of this thesis while allowing me maximum time with my family. Your knowledge of the related literature, statistical techniques, and proper methodology was invaluable to me.

Portions of these data were presented at the Human Factors and Ergonomics Society 51st Annual Meeting (Baltimore, Maryland, 2007) and appear in the proceedings of that conference (Wilkison, Fisk, & Rogers, 2007). This research was supported in part by a grant from the National Institutes of Health (National Institute on Aging) Grant P01 AG17211 under the auspices of the Center for Research and Education on Aging and Technology Enhancement (CREATE). This research was also supported in part by contributions from Deere & Company. I thank Jerry Duncan for his support and advice on this research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

MM	Mental Model
MCP	Multiple Comparison Procedures
HSD	Honestly Significant Differences
SA	Situational Awareness

SUMMARY

As the tasks humans perform become more complicated and the technology manufactured to support those tasks becomes more assistive and adaptive, the relationship between humans and automation transforms into a collaborative system. In this system each team member depends on the input of the other to reach a predetermined goal beneficial to both parties. Studying the human/automation dynamic as a social team provides a new set of variables affecting performance and effectiveness previously unstudied by automation researchers. One such variable is the shared mental model (Mathieu, Heffner, Goodwin, Salas, & Cannon-Bowers, 2000).

This study examined the relationship between mental model quality and collaborative system performance within the domain of a navigation task. Participants navigated through the confines of a simulated city with the help of a navigational system that performed at two levels of accuracy; 70% and 100%. Participants with robust (accurate) mental models of the task environment identified automation errors when they occurred and optimally navigated to specified destinations. Conversely, users with vague (inaccurate) mental models of the task environment were less likely to identify automation errors, and chose inefficient routes to specified destinations. Thus, mental model quality proved to be an efficient predictor of navigation performance. Additionally, participants with no mental model performed as well as participants with vague mental models. The difference in performance between users with no mental models and users with inaccurate mental models was the number and type of errors committed.

This research is important as it supports previous assertions that humans and automated systems can work as teammates and perform teamwork (Nass, Fog, & Moon, 2000). Thus, other variables found to impact human/human team performance might also affect human/automation team performance just as this study explored the effects of a primarily human/human team performance variable, the mental model. Additionally, this research suggests that a training program creating a weak, inaccurate, or incomplete mental model in the user is equivalent to no training program in terms of pure performance. Finally, through a qualitative model, this study proposes mental model quality affects the constructs of user self confidence and trust in automation. These two constructs are thought to ultimately determine automation usage (Lee & Moray, 1994). To validate the model a follow on study is proposed to measure automation usage as mental model quality changes.

CHAPTER 1

INTRODUCTION

Automation has seeped into almost every facet of human life and its proliferation will surely continue with the advent of new and better technologies. As the tasks humans perform become more complicated and the technology manufactured to support those tasks becomes more assistive and adaptive, a collaborative system consisting of a human user and an automated aid is formed. In this system each team member depends on the input of the other to reach a predetermined goal beneficial to both parties.

Much of the automation research in the human factors field has studied automation in the context of a decision aid (see Parasuraman & Riley, 1997; Parasuraman, Sheridan, & Wickens, 2000, for reviews). While scientifically beneficial, this research may have excluded some variables affecting the performance of the human/automation team. Thinking of a collaborative system as a team and studying this dynamic in the same manner as human team researchers, other variables may be uncovered that will provide valuable insight into the future of automation design and human use. To lay the foundation for this idea a review of four separate, but related research areas follows: 1) automation 2) computers as social actors 3) team performance and 4) situational awareness.

Automation Research Review

A simple definition of automation is the execution of a function by a machine that was previously carried out by a human (Parasuraman & Riley). Sheridan (2002) explained further the wide and exhaustive categories of automation:

> Automation refers to (a) the mechanization and integration of the sensing of environmental variables (by artificial sensors); (b) data processing and decision

> making (by computers); and (c) mechanical action (by motors or devices that apply forces on the environment) or "information action" by communication of processed information to people. It can refer to open-loop operation on the environment or closed-loop control."(p. 9)

By these definitions automation could be as simple as a stapler or as complicated as a Combat Information Center aboard a Navy Destroyer. With such an all encompassing definition, "automation can highlight, alert, filter, interpret, decide, and act for the operator."(Lee, 2005, p.1577) Sanchez (2005) identified six definitions within the automation literature and proposed his own, insightful definition that combined the most relevant aspects from each. For the purposes of this paper I will use his definition: "automation is a technologically-based system used to partially or fully assist the human in tasks involving sensing, detecting, processing information, making decisions and/or executing actions."(p. 11)

Even with this more precise definition of automation, the number and types of different systems and tasks within this problem space can seem daunting. To standardize and simplify the domain even further researchers utilize a taxonomy developed by Parasuraman et al. (2000). They proposed four different types of automation based on its role: 1) information gathering, 2) information analysis, 3) decision selection, and 4) action based (See Figure 1). These four types of automation are similar to a basic problem solving strategy where one would gather information about a problem, determine its relevance, come to a decision, and then see it through to some outcome.

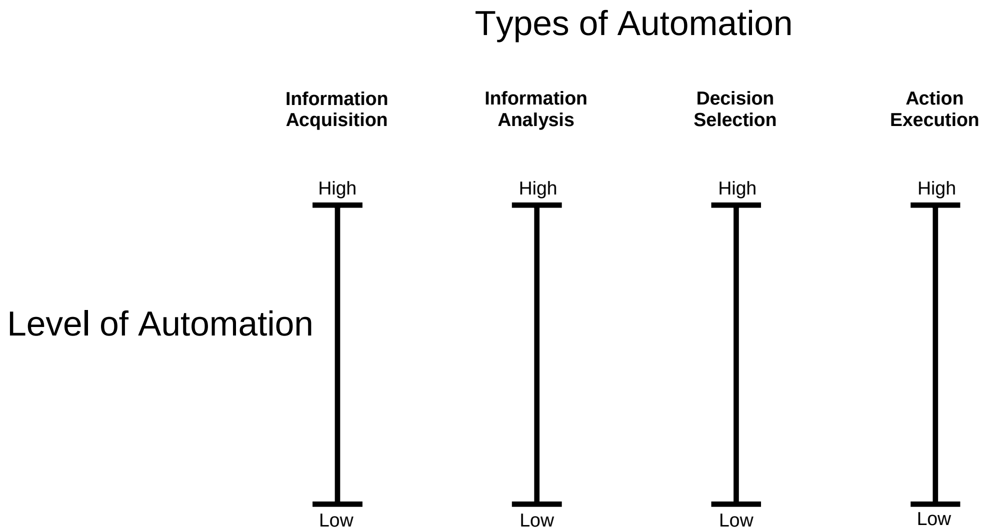

Figure 1. Types of automation with level of human control or lack of control. Adapted from Parasuraman, Sheridan, & Wickens (2000).

Parasuraman et al. (2000) also categorized automation into ten different levels based on the automation's level of control or human lack of control. For example, automation in the first takes action only when directed by the human whereas automation in level ten acts autonomously without human supervisory control. An automated system could potentially act within all four types with varying degrees of level at each type. For example, a global positioning system (GPS) navigational aid acts at a very high level of automation during the information gathering stage. It receives signals from positioning satellites without permission of its human user to determine the systems current location. During the information analysis stage, the GPS operates at a middle level as it takes the user-provided destination and compares it with the current location. In the decision selection stage the GPS offers a recommended solution (high level) but does not actually make the decision (low level) on how to proceed. During the action implementation stage the human makes all movement towards the desired destination, but with periodic checks with the GPS (low level).

Where do collaborative systems fit into the automation levels and types proposed by Parasuraman et al. (2000)? Collaborative system automation could potentially perform within all four classifications of automation (information acquisition and analysis, decision selection, and action implementation). It can gather information, process it, make a recommendation, and even execute actions if the human teammate so desires. On the other hand, collaborative system automation should not perform duties in all 10 levels classified by Parasuraman et al. Any level that has the automation acting without regard to the human teammate is not collaborative. Collaborative automation works with the human towards a common goal and constantly gathers and processes new information while updating its recommendation and waiting for a final human decision.

There are many factors that may affect the human/automation relationship and the resulting performance of the team. Factors include mental workload, skill, confidence, task complexity, fatigue, and risk to name a few. These factors either directly or indirectly affect user trust, reliance, and compliance with the automated system. Inappropriate levels of trust are manifested in two types of errors 1) disuse, which is the user's under reliance of the automation, or not using the automation as frequently as to benefit from its fullest capability and 2) misuse, which is over-reliance on automation, or using the automation when it performs incorrectly (Parasuraman & Riley, 1997).

Several studies have examined how disuse and misuse errors, within the context of signal detection theory, adversely affect performance. Dzindolet, Peterson, Pomranky, Pierce, and Beck (2003) found that errors on the part of an automated decision aid degraded user trust and influenced compliance with automated aids. More specifically, when subjects observed the automation making an easily detectable error, they tended not to use that automation even when it was accurate in its future suggestions (disuse). Dixon, Wickens, and McCarley (2006) found the type of error committed by the automation affected user compliance and reliance; false alarms being easier to detect than misses tended to degrade user trust more than misses when they occurred (disuse). Wickens, Dixon, and Johnson (2006) showed how error rate and task priority also affected user performance (misuse and disuse). Madhavan and Wiegmann (2005) found that labeling an automated aid as an expert or novice also influenced a user's trust and would therefore influence use and performance (misuse). Cost of an automation error is yet another factor that has been shown to influence user trust in automated systems. As the cost of making an error rises, humans tend to rely less on automation and take more time and effort to confirm any decision recommended by the automation (misuse; Ezer, 2006). Obviously, there is a host of factors and variables that cause misuse and disuse, and correspondingly affect user trust in and reliance of automated systems. However, looking at this relationship as a collaborative system uncovers even more possible variables.

Automation as a Teammate and Social Actor

Within the collaborative system the human and the automation work as a team, but can humans actually view automated systems as teammates and equal partners in work? Numerous studies have shown evidence that humans treat automation and specifically computers as social actors even when there is no reason to do so **(**Klein, Woods, Bradshaw, Hoffman, & Feltovich, 2004; Nass & Moon, 2000; Nass, Fogg, & Moon, 1996; Nass, Moon, & Carney, 1999; Nass, Moon, & Green, 1997; Nass, Steuer, & Tauber, 1994; Reeves & Nass, 1996). For example humans, exposed to a limited amount of cues, may apply social attributes to computers such as politeness, stereotypes, and notions of 'self' and 'other' (Nass et al., 1994). Researchers have also shown that by merely putting a colored arm band on a user and a matching colored border around a computer monitor, users will view the computer as a teammate and refer to it as such (Nass & Moon). In all of these studies the participants realized they were interacting with the computer and not some programmer behind the automation. However, they still unknowingly interacted with these machines in a social way, even applying gender roles in one circumstance with the subtlest of cues (Nass et al., 1997).

Although humans may be able to view computers and automation as social actors, can these same automated systems be considered teammates and perform teamwork? Interdependence and team identity play an important role in the team interaction between humans and automation. Nass et al. (1996) manipulated the identity (told participants that they were part of a team with a computer) and interdependence (performance was based on combined efforts of the human/computer team) of a human-computer team, and found that humans viewed themselves as teammates with the computer. They were more likely to comply with the computer's advice, found the computer to be friendlier, and perceived the computer's information to be of a higher quality. Thus, telling users their performance will be evaluated in combination with their automated "teammate" should generate team identity and interdependence and lead to teamwork. Salas, Kosarzcki, Tannenbaum, and Carnegie (2005) defined teamwork as a "set of interrelated behaviors, actions, and decisions that yield a shared and valued outcome."(p. 136). Researchers have shown people can and do perceive behavioral traits from automation (Nass et al., 1994). Technology has given automation the ability to take independent action or work in consort with users. Decisions recommended by automated systems are often in pursuit of the human defined goal. Therefore, collaborative systems can be thought of as teams and studied as such.

Knowing humans can view automation as a teammate and perhaps unknowingly apply team attributes to the automation, a set of variables previously unstudied may become important. In other words, any factor or variable that has been shown to affect the performance of human teams could potentially affect the performance of collaborative system teams in similar ways.

One variable that repeatedly surfaces in the team studies literature is the idea of a shared mental model. Researching the effects of mental models on collaborative system performance is supported by Langan-Fox, Anglim, and Wilson (2004), who suggested that team research findings concerning mental models could potentially be applied to joint cognitive systems of humans and computers.

Before examining the effects of mental models on collaborative systems a more detailed review of the concept is needed. Mathieu, Heffner, Goodwin, Salas and Cannon-Bowers (2000) stated, “Mental models are organized knowledge structures that allow individuals to interact with their environment.”(p.274). Whereas the idea of humans possessing and using mental models may seem intuitive, there exists some disagreement in the mental model literature as to the definition and importance of this construct (Langan-Fox et al., 2004; Wilson & Rutherford, 1989). The most relevant definition of mental models for the purposes of this study and collaborative systems in general comes from Langan-Fox et al.; “Mental models are simulations that are run to produce qualitative and quantitative inferences, underpin our understanding of a system, and allow us to describe, predict, and explain behavior of a system.”(p 334).

Particularly important to teams and teamwork, shared mental models allow for a common understanding between group members of the task, the environment, and the abilities and responsibilities of each other (Cannon-Bowers & Salas, 1998). Salas, Prince, Baker, and Shrestha (1995) stated teams use mental models in two ways: 1) when communication is prohibited or degraded, shared mental models allow team members to anticipate teammate actions and their requirements for information or assistance, and 2) the shared mental model allows teammates to operate from a common frame of reference where everyone acts under the same perceived facts and assumptions. Previous research has shown that well performing teams have robust, shared mental models whereas underachieving teams have inaccurate or vague shared mental models (Mathieu, Heffner, Goodwin, Cannon-Bowers, & Salas, 2005; Mathieu, et al., 2000; Serfaty, Entin, & Johnston, 1998). The question then becomes, given that humans and automation can form a collaborative system, does performance improve or degrade with the quality of the mental model established in the human and programmed into the automation?

Mental Models and Situational Awareness

Not only do mental models support better team performance, but several situational awareness (SA) researchers believe that it plays a pivotal role in the formation of SA. SA defined by Endsley (1995), "is the perception of elements in the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future."(p. 36). Endsley parsed SA into three levels that simplify the construct, but that also point out the influence of mental models on the quality of SA. Level 1 entails the perception of the environment and the task space. Level 2 encompasses the comprehension of the environmental data and building of appropriate mental models. Level 3 is the use of those mental models in the prediction of future states of the environment and task domain. While each level of SA is individually important, without the construction of appropriate mental models, human ability to use SA to forecast future states is degraded and will ultimately affect task performance and problem resolution. Dominguez (1994) also expressed the importance of mental models when she defined SA as the "continuous extraction of environmental information, integration of this information with previous knowledge to form a coherent mental picture (model), and the use of that picture in directing further perception and anticipating future events."(p. 11). Dominguez viewed SA as having two dimensions; the process and the product. More specifically she holds that throughout the research literature, the product dimension of SA is commonly referred to as mental models. Judging from the work of Endsley and Dominguez, mental models are of great importance in the construction and utilization of SA in the individual.

Team SA is also greatly dependent on shared mental models. Aircraft and submarine crews, nuclear power and manufacturing plant staffs, and fire fighters are examples where team SA is important in not only completing the mission, but in ensuring everyone survives. Endsley (1995) found that higher level SA in teams was developed by the sharing of mental models that made communication more efficient and led to higher levels of performance. Thus, it would appear that the increase in team performance due to shared mental models (Salas et al., 1995; Volpe, Cannon-Bowers, & Salas, 1996; Stout, Cannon-Bowers, Salas, & Milanovich, 1999) was due in part to the establishment of SA with the help of those mental models. The focus of this study is the establishment of shared mental models within the collaborative system team, the extension of that mental model to a navigational task, and the group performance evaluation when mental model quality is manipulated.

Current Study and Hypotheses

In this study mental model quality was the independent variable in a split plot design. The quality of the participant mental model was defined operationally by their performance on a series of acquisition tasks. A poor or vague mental model group (low acquisition group), an excellent or robust mental model group (high acquisition group), and a no mental model group (no acquisition group) performed a navigation task within the confines of a simulated city with the help of an automated navigational aid performing at two levels of accuracy. The human and the aid formed a collaborative system with the goal of navigating to specified destinations from predetermined start points as efficiently as possible.

To establish the quality of the mental model within the three groups a training method consisting of city relational, directional, and construction tasks was used. Participants spent time learning the spatial relationships of the city within an allotted amount of time. After study they continued to establish, to varying degrees, their mental models of the task environment by answering questions about the spatial relationship of the city blocks. Relational questions featured questions about city blocks adjacent to one another. Directional questions addressed block distance and direction from other blocks. Finally, a map reconstruction task assessed the participants' ability to re-build the map from a random assortment of city blocks. Participant performance on these tasks operationally defined the quality of their mental models. Participants in the high acquisition group were required to answer 90-100% at the conclusion of five iterations of the acquisitions tasks. Low acquisition group participants were not required to reach a specified criterion, but were only allowed to complete the acquisition iteration one time. Lastly, the no acquisition group had no prior experience with the city and performed the navigation tasks only.

This method of building different levels of mental model quality in participants is not only supported by Gilbert and Rogers (1999), but also by the work of Thorndyke and his colleagues (Goldin & Thorndyke, 1982; Thorndyke & Stasz, 1980; Thorndyke & Hayes-Roth, 1982). Thorndyke, through a series of studies, examined how humans acquire spatial knowledge from maps and navigation and how they use that knowledge to make various estimates and decisions. He suggested that people create mental models to learn maps and then use those models to make spatial judgments. He also contended that with greater exposure to the map and experience with the spatial arrangements, people would increase the accuracy of their judgments. Certain procedures were found to be beneficial to acquiring spatial knowledge from a map. They were partitioning, imagery, and relation encoding. Partitioning refers to the procedure whereby a participant would learn only a small segment of the map before moving on to a new, unlearned portion. The proposed method supports this by forcing participants to attend to a specific portion of the map at one time (relational questions). Imagery was the procedure in which participants attempted to place the map information into a mental model to be retrieved later. Relation encoding was the procedure that entailed participants attending to the actual directional relationships of prominent features on the map. They noticed that a group of buildings were below a river or that a certain building was east of another. This procedure is reflected in the directional and relational questions used in the current method. In summary, this study's method of building mental models of map data builds upon those procedures shown to support the acquisition of spatial information from maps.

During the navigation test phase, participants navigated from predetermined start points on the map to destinations. Since only the origin and destinations were provided along with a blank map, they were forced to rely on their mental models, their automated teammate, or a combination of the two. There were two levels of automation support: 1) 70% automation accuracy, and 2) 100% automation accuracy. The 70% level was chosen based on previous research that indicated 70% automation accuracy allowed participants to notice a decrement in support, but the accuracy level also maintained some trust and participants still relied on the aid to a certain extent (Kantowitz, Hanowski, & Kantowitz, 1997; McGuirl & Sarter, 2006). After completing the trial blocks participants engaged in a test of learning consisting of all the acquisition tasks from early in the procedure (map reconstruction, relational and directional questions) to measure improvement or decline in mental model quality. Critical measures were the number of optimal routes selected, the number of misuse and disuse errors committed, and subjective assessments of automation reliability after each block of trials. At the conclusion of the study, participants completed a subjective survey to measure their self-confidence and the level of trust in the automation

This method was designed to answer the following research questions. **RQ1**. Does the quality of user mental model positively affect performance in a collaborative system? **H1**. With the assistance of an automated teammate, participants in the high acquisition group should perform better than participants in the low and no acquisition groups. Participants with a detailed mental model will more accurately identify errors in the automation and will optimally rely on the aid. Stated another way, users in the high acquisition group will rely on the automation when it is correct and will determine their own solution when the automation is incorrect, thus optimally performing.

RQ2. Is having a vague mental model worse than having no mental model at all? **H2**. Participants in the no acquisition group will have no experience with the city and no resources to draw upon when navigating the city except for the automated navigational aid. They may be forced to blindly trust the automation and its recommendation. On the other hand, participants in the low acquisition group should have a vague and most likely inaccurate mental model to draw upon when navigating. This may lead them to distrust the automation when it is in their best interest to comply with its recommendations.

CHAPTER 2

METHOD

Participants

29 male (age M = 20.38, SD = 1.43) and 31 female (age M = 20.00, SD = 1.57) undergraduate students from the Georgia Institute of Technology participated in this study. They were enrolled in an introductory psychology course and received credit or monetary compensation for their participation. Participants were required to possess 20/40 vision (corrected) or better to read the displays and distinguish between items within the task trials. Red/green colorblindness was a disqualifier for the study due to the use of color as a discriminator among common shapes within the task trials. Figure 2 shows the ethnicity of participants by percentage and shows the diversity of the sample population.

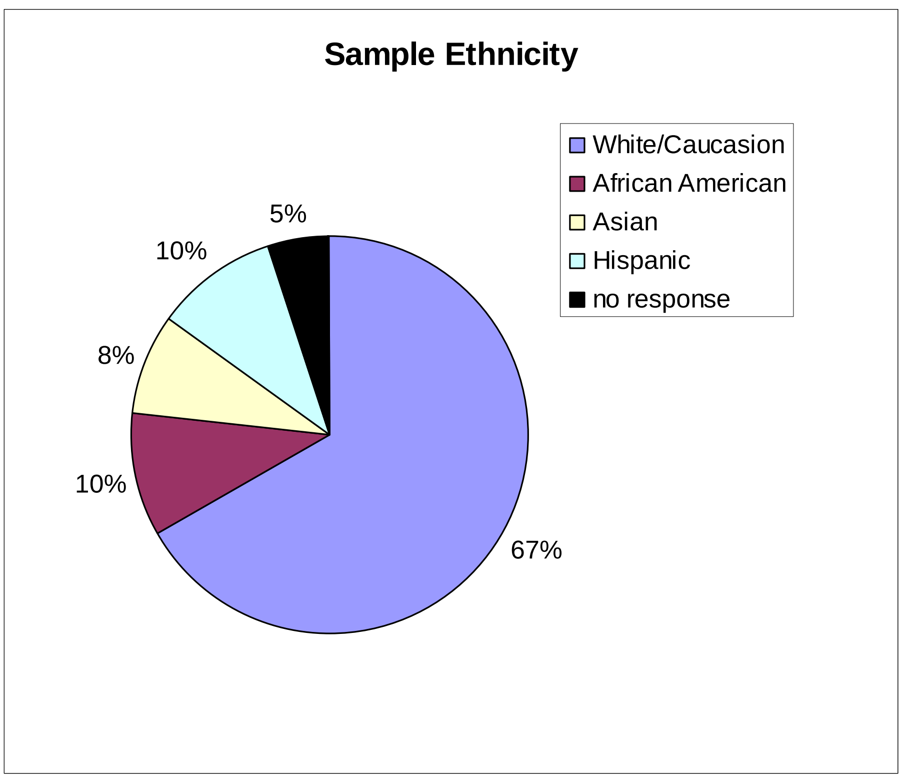

Figure 2. Sample ethnicity by percentage.

Materials

Prior to beginning the experimental tasks and the acquisition of mental models participants conducted cognitive and perceptual abilities tests. The Digit Symbol Substitution Test (Wechsler, 1997), Reverse Digit Span Test (Wechsler), and the Shipley Vocabulary Test (Shipley, 1986) measured perceptual speed, memory span, and verbal ability respectively. Two tests were used to measure spatial ability, the Cube Comparison Test and the Paper Folding Test (Ekstrom, French, & Harmon, 1979). Ability tests were used to describe the participant sample and compare the three experimental groups at the conclusion of the study. Participants also had their vision tested to the 20/40 level using the Reduced Snellen chart (from 14 inches) and the Snellen chart (from 20 feet) and completed Ishihara's test for color deficiency (Ishihara, 1960). A demographics and health survey was administered to assist in describing the study sample. A trust in automation survey (Jian, Bisantz, & Drury, 2000; see appendix A) was also administered to gauge the preconceived notions of automation reliability within the sample.

A map (see Figure 3) similar to the one used by Jastrzembski, Roring, and Charness (2006) constructed of streets and buildings was the task environment. The city contained a four by four block structure graphically simulated. Each city block was unique by a combination of shapes, colors, and orientation. All stimuli and tasks were presented on a 15 inch, color monitor. The experiment was programmed using Java. Participants used an optical mouse and a standard QWERTY keyboard for input devices. A nine question subjective survey (see appendix B) using a 5 point Likert Scale was presented to participants after the completion of all experimental tasks to determine participant confidence and trust in their automated teammate.

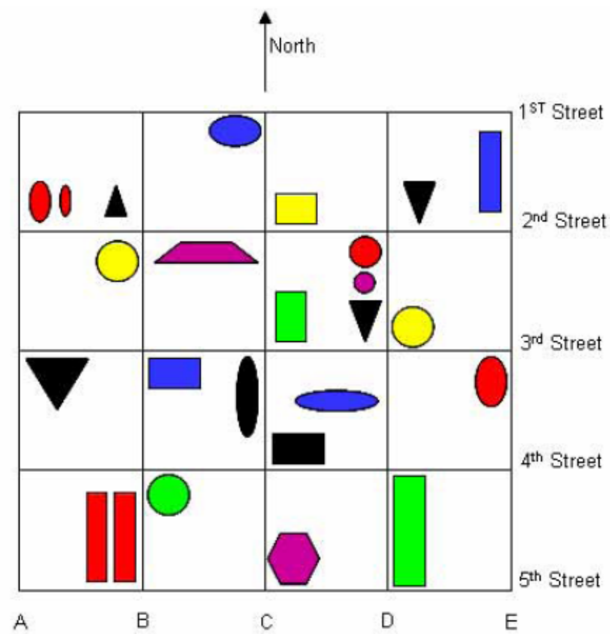

Figure 3. Task environment; simulated city as seen by participants in map study.

Design

A 3x2 split plot design was used with three levels of acquisition (between subjects); no, low, and high, crossed with two levels of automation reliability (within subjects); 70% and 100%. The study manipulated the acquisition level of the participants by varying their exposure to and experience with the simulated city map. The map reconstruction task along with relational and directional questions made up one acquisition iteration. The high acquisition group completed five iterations of the acquisition tasks before moving on to navigation trials. The low acquisition group completed one iteration of the tasks, while the no acquisition group was restricted from viewing the map prior to navigation trials.

Automation reliability during navigational trials was determined by the accuracy of the automation recommendation. During the 70% reliability condition, the automation provided the correct solution on 14 out of 20 trials. The automation performed without error during the 100% reliability condition. The order of the 70% and 100% trial blocks was counterbalanced across participants to guard against recency effects (Sanchez, Ezer, Rogers, & Fisk, 2006). Critical measures during the navigation task trials were the number of optimal routes selected, along with the number of misuse and disuse errors committed. Between navigation trial blocks, participants answered one subjective question concerning the accuracy of the automation for the preceding trials. Once the trial blocks were over, participants performed a test of learning consisting of the acquisition tasks to measure learning acquired during navigation trials. At the conclusion of the study participants completed a subjective survey designed to measure self-confidence and trust in the automation.

Tasks and Procedure

Once participants arrived they completed an informed consent form, demographics and health survey, trust in automation survey, and the battery of ability tests described earlier. After completing the ability tests, participants began the acquisition phase of the study. See Appendix C for examples of trials and Appendix D for detailed instructions. The method used in the Gilbert and Rogers (1999) study served to build different levels of city mental models in the low and high acquisition groups. To familiarize participants with the program and the input devices they completed an introductory task (see Appendix C, Figure 19). They performed three trials where they moved shapes from the peripheral edge of the screen to assigned blocks on a four by four city grid. On each trial there were four numbered blocks; these numbers corresponded to a number on the four by four grid map. Once the participants dragged and dropped the blocks onto the appropriate grid they selected a submit button with the mouse. After selecting the submit button, they received feedback on the block placement task and moved onto the next practice trial. Participants completed three trials regardless of performance.

Acquisition Phase

In the mental model acquisition phase participants studied the city map for one minute, performed a distracter task (subtract a value from a given value for 30 seconds), and then performed tasks to show they understood the spatial relationship among the 16 city blocks. Acquisition was tested with a series of relational and directional questions along with a map reconstruction task. The map reconstruction task (see Figure 4) entailed showing the entire city separated by blocks and displayed on the periphery of the 15 inch screen. Participants moved, via the mouse, the blocks onto their appropriate grids on a checkerboard reference system. Once the blocks were in position, the participant selected the submit button with the mouse and received feedback in the following form: “you got 14 out of 16 correct”.

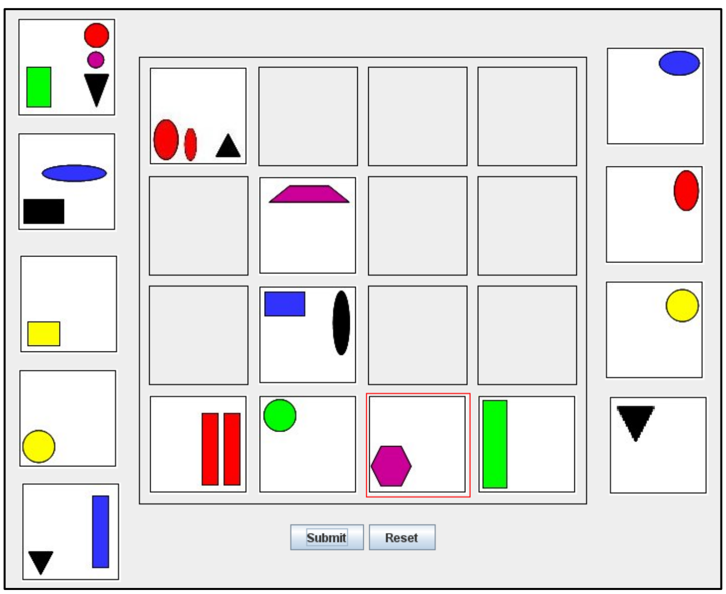

Figure 4. Map reconstruction task.

Relational questions (see Figure 5) presented one city block and then asked in the form of a True/False question if the block was north, south, east or west of the second city block. These questions used blocks that were in direct contact with each other. For each relational block of trials, there were two practice questions and 16 test questions. These questions were randomly selected from a bank of 36 trials. After submitting the true/false answer, the participant received feedback in the following format: “Sorry you answered incorrectly” or “That is correct”.

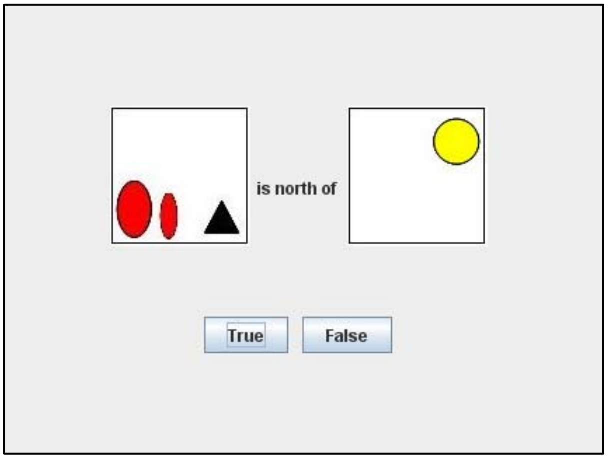

Figure 5. Example of relational question.

The directional questions (see Figure 6) displayed two city blocks and asked how many blocks and in what direction one would travel to get from one block to the other. Directional questions involved blocks that were separated by up to three blocks. Two practice questions were administered before the 16 test questions. These questions were also randomly selected from a bank of 36 trials. Feedback after each trial was in the following format: “Sorry you answered incorrectly” or “That is correct”.

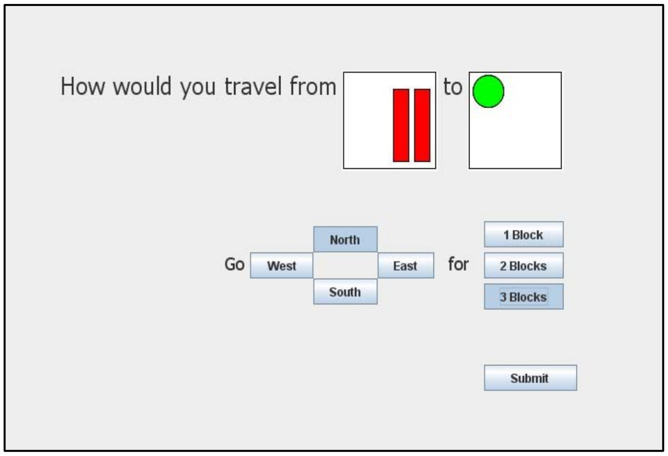

Figure 6. Example of directional question.

The high acquisition group completed five iterations of the acquisition trials (map reconstruction, relational questions, and directional questions) before moving on to navigation trials while the low acquisition group completed one iteration before moving on. The no acquisition group skipped the acquisition phase of the study and went directly into the navigation phase.

Navigation Phase

The navigation phase consisted of two blocks of 20 trials each. The 70% automation support and a 100% automation support block were counterbalanced across participants. The six error trials within the 70% block were randomly placed for each participant with the following heuristic: error trials could be placed in trials 1-20 at random, but no more than two error trials could occur simultaneously. For instance a random selection of error trials could have placed one error trial in trial 1 and another in trial 2, this would enact the heuristic within the program and trial number 3 would automatically be filled with a correct automation trial. The 14 correct trials in the 70% automation support block and the 20 trials in the 100% automation support block were randomly selected from a pool of 64 trials. The 64 trials were constructed to ensure each of the 16 city blocks was utilized as the origin in at least 4 trials.

Each navigation trial (see Appendix C, Figure 24) displayed a blank city grid (the map with all colored shapes removed), an origin, a destination, and an automated recommendation. The origin was placed in the upper left side of the screen and consisted of one city block with a red, oval icon placed on one side of the block. The icon indicated the exact, current location of the origin in relation to the center of the block. For instance, if the icon was placed on the top side of the block, then the participant would know that the origin was on the north side of that particular block. The destination was placed just below the origin and took the form of another city block with a red icon indicating the exact location of the destination. The automated route recommendation was depicted as a green line placed on the blank city map. At one end of the green line was a red, oval icon. Participants were told the red, oval icon connected with the green line corresponded to the origin which was also given in the upper left corner of the screen. Participants used the mouse to drag the red icon from the given destination to the given origin. Again, in the 70% block trial only, there were six trials in which the green line (automation's recommendation) was inaccurate. In all other trials the green line correctly depicted an optimal route from origin to destination.

Once participants moved the icon to a desired location on the map they had the choice of submitting that route for feedback or resetting the map. Resetting the map erased the route traced by the participant, and placed the icon back to its original location (origin). The reset option was only available once per trial. The submit option would result in one of three feedback conditions, 1) Congratulations! You found the optimal route. 2) Good job, you found your way and 3) Sorry, you ended up in the wrong place. Feedback condition 1 resulted when the participant moved the icon from origin to destination in the shortest distance. Feedback condition 2 resulted when the participant moved the icon from origin to destination, but took unnecessary additional turns. Feedback condition 3 resulted when the participant placed the icon anywhere on the map but the destination.

After each trial block, participants were cued with the following statement; "Please indicate, using a number, the reliability of the automated navigation system over the last block of trials. (Example: I think the navigation system was XX% reliable.)" Using the keyboard, participants entered a two or three digit number. Once the entry was made participants moved onto the next block of trials or the test of learning.

Test of Learning

Upon completion of the navigation phase, participants conducted the map reconstruction task, 16 trials of relational questions, and 16 trials of directional questions. Only members of the no acquisition group received practice trials during the test of learning. Once the test of learning was complete the experimenter instructed the participants to complete the 9 question subjective survey. When this was complete the participants were debriefed and released.

CHAPTER 3

RESULTS

Unless stated otherwise, alpha was set at .05 for all statistical tests, all t tests utilized the two tailed analysis, and all error bars are standard error bars.

Abilities Tests

To explore group differences on the five abilities tests a MANOVA was used, Wilk's λ = .887, p = .710. Table 1 depicts the univariate ANOVAs for each statistical test. These results indicate that the three acquisition groups did not significantly differ in their performance on the five abilities tests.

Table 1. Analysis of Variance for Abilities Tests

Ability Test	F	p
Cube Comparison	0.827	0.442
Paper Folding	1.003	0.373
Reverse Digit Span	1.622	0.207
Shipley Institute of Living Scale	0.329	0.721
Digit Symbol Substitution	1.187	0.313

Acquisition Performance

The low and high acquisition groups performed one and five iterations of the acquisition tasks respectively. The low acquisition group's performance was marginal on their iteration (M = 33.03%, SD = 11.80%). This measure provides support for the assertion that members of this group entered into the navigation task with a weak, vague, or incomplete mental model. The high acquisition group's performance on the five iterations is depicted in Figure 7. The mean performance for members of this group on the last iteration before the navigation task was near ceiling (M=96.26%, SD= 3.72%). This performance supports the idea that the high acquisition group entered into navigation task with a robust and very accurate mental model of the city map.

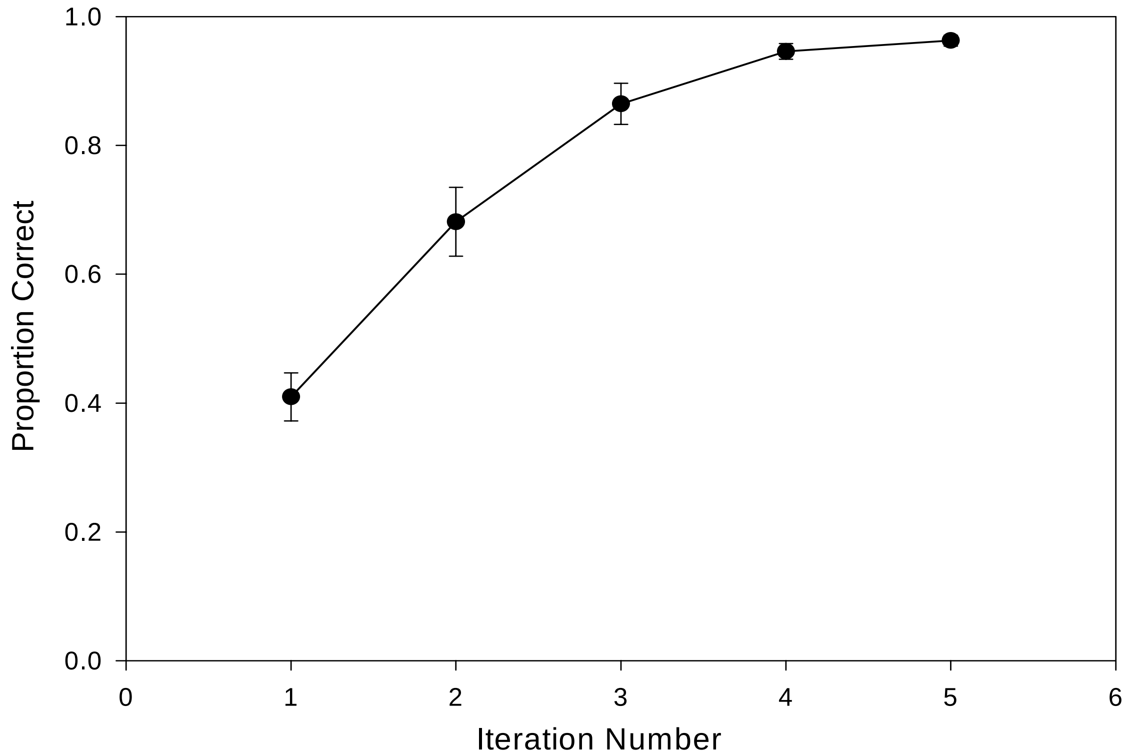

Figure 7. High acquisition group performance on each of the five acquisition iterations prior to navigation task.

Navigation Task

After acquiring varying levels of experience (and presumably levels of mental model) with the task environment, the three acquisition groups conducted two blocks of 20 navigation trials. Among the measures collected during navigation, four were particularly relevant; optimal routes selected, misuse errors, disuse errors, and response time. To analyze the first measure, a 3x2 (acquisition group x automation support level) split plot ANOVA was conducted. There was a main effect for Acquisition Group, $F(2, 57) = 44.41$, $p<.0001$; a main effect for Automation Support, $F(1, 57) = 191.361$, $p<.0001$; and an interaction for Group x Support, $F(2, 57) = 29.40$, $p<.0001$. Figures 8 and 9 depict the means for the main effect of group and interaction effect respectively for the dependent measure of optimal routes selected.

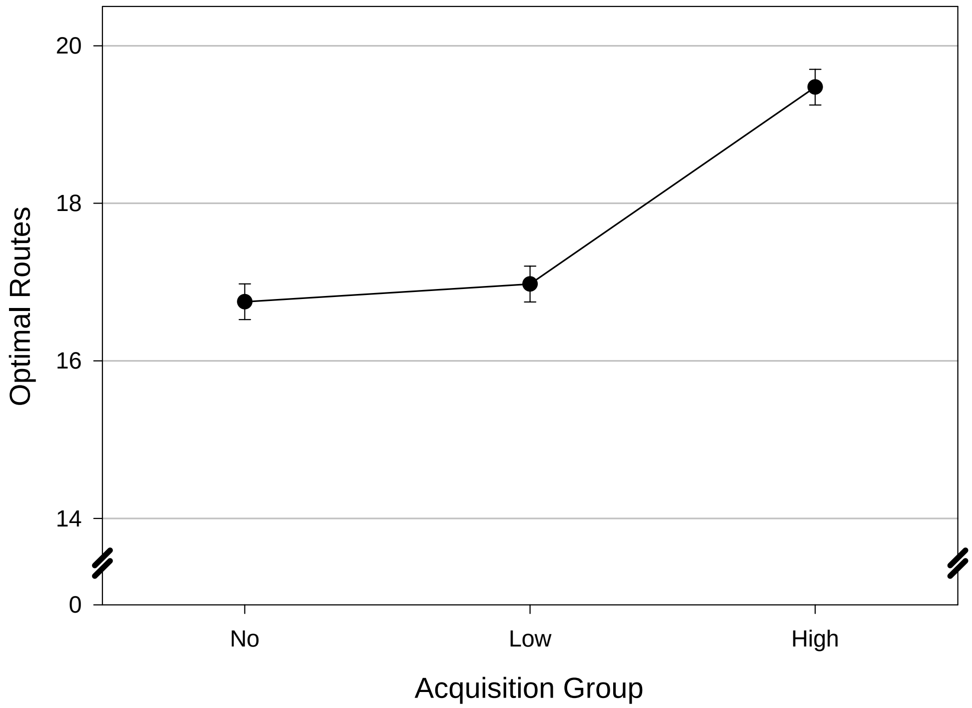

Figure 8. Means of optimal routes selected by acquisition group during both blocks of navigation trials.

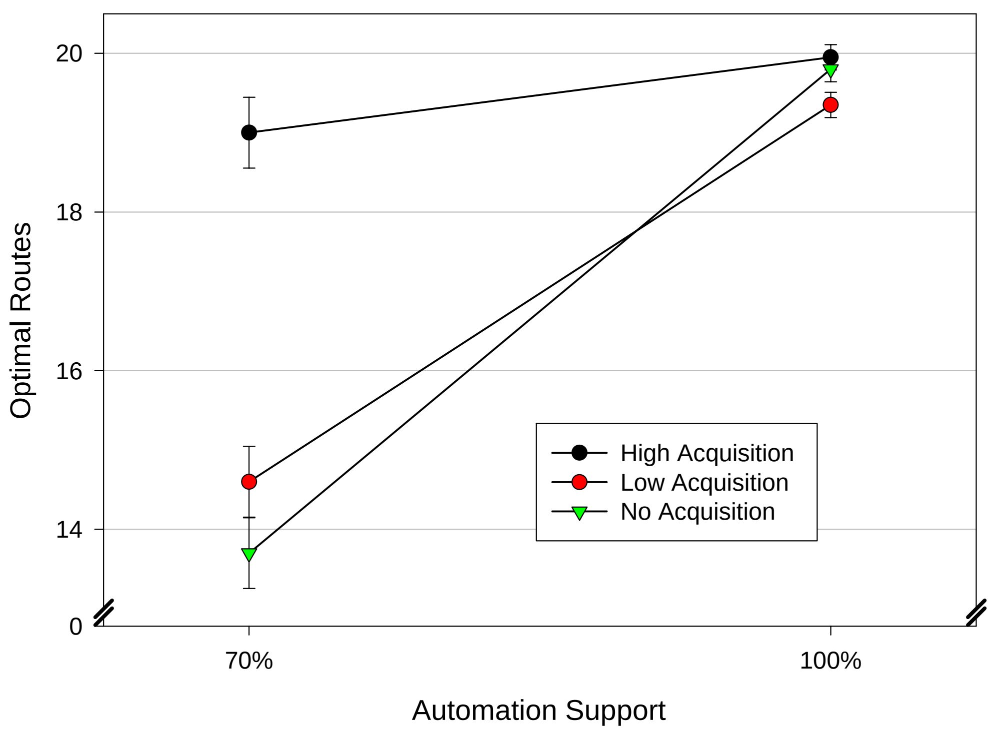

Figure 9. Marginal means of optimal routes plotted by level of automation support.

To explore the interaction, independent sample t-tests were conducted. In this analysis the acquisition group performances during the 70% automation accuracy block and the 100% automation accuracy block were tested separately. Using Bonferroni's correction, Alpha was set to .0083. Table 2 shows the results of this analysis. These results indicate that high acquisition group performance during the 70% block trial was significantly better than the performance of both the no and low acquisition groups. Additionally, the no and low acquisition group performances during this block of trials were not statistically different. During the 100% block of trials all three groups performed at the same level, with no statistical difference between them.

Table 2. Post hoc independent groups t-tests for split plot ANOVA of optimal routes

Trial Block	Groups	t	p
70% Automation Accuracy	No vs Low	-1.24	.224
	No vs High	-15.20	.0001*
	Low vs High	-5.96	.0001*
100% Automation Accuracy	No vs Low	1.67	.105
	No vs High	-1.44	.159
	Low vs High	-2.32	.026

Note. * Sig. at alpha = .0083

Response Time

To measure response time the program incorporated a time log that began immediately upon trial initiation. The time log stopped when participants selected the SUBMIT button with the mouse. A split plot ANOVA found a main effect of automation support, but no effect for acquisition group and no interaction. Table 3 contains the ANOVA summary table and Figure 10 shows the means of response time for acquisition group versus automation support.

Table 3. Analysis of Variance for Response Time

Source	df	F	p
Group (G)	2	2.994	0.058
between-group error	57	(17244151.160)	
Automation Support (A)	1	13.376	0.001*
G x A	2	0.247	0.782
within-group error	57	(9987680.849)	

Note. Values enclosed by parentheses are mean square errors. *p<.01.

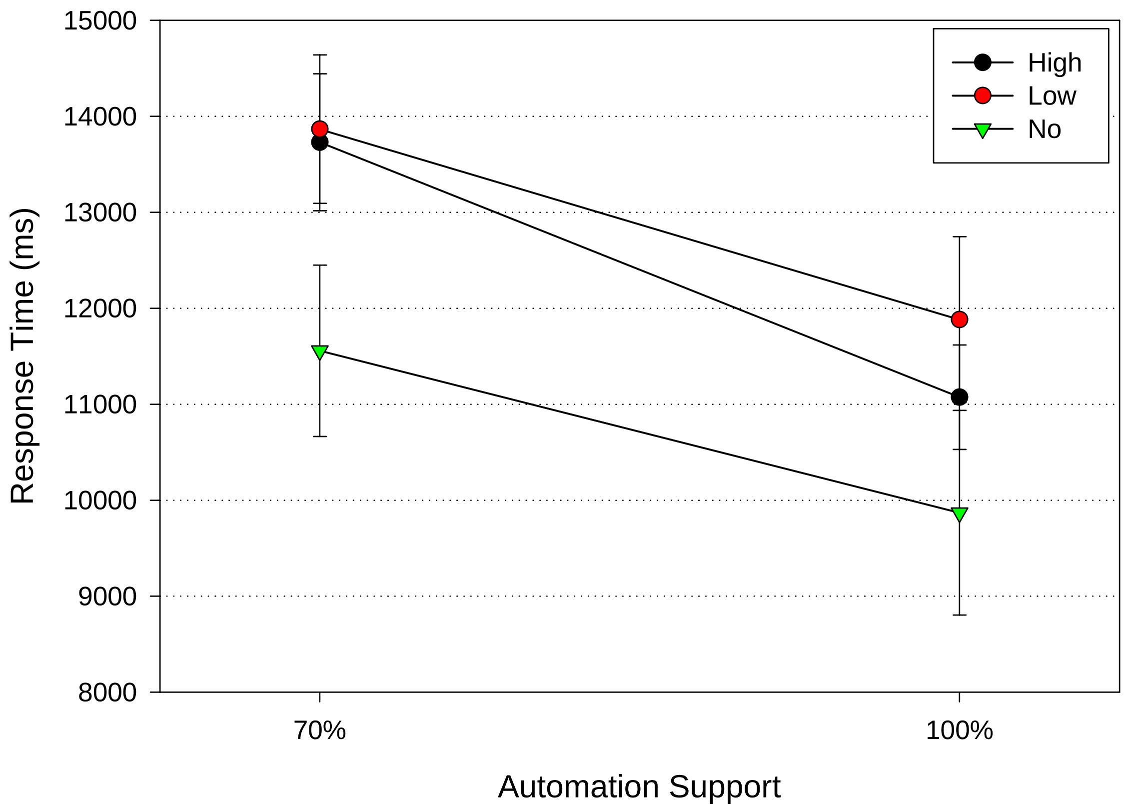

Figure 10. Means of response time for acquisition group versus automation support.

Table 3 and Figure 10 show that overall, participants completed trials within the 100% block faster than in the 70% block. Additionally, within each trial block groups completed trials at equal time intervals. Put another way mental model quality did not influence response time significantly, but automation support did.

Errors

While there was no significant difference in performance during the 100% block of trials there were some trends among the groups worth noting. Table 4 shows the number of participants committing disuse errors (misuse errors were not possible as the automation's recommendation was always correct) within the 100% block, the number of disuse errors committed, and the trial number in which the error occurred.

Table 4. 100% trial block trends.

participant ID	acquisition group	disuse errors	trial number
527	High	1	7
107	Low	1	3
108	Low	1	3
110	Low	1	20
112	Low	2	4 & 8
117	Low	1	16
124	Low	4	1,2,6, & 9
125	Low	3	9, 10, & 12
10	No	1	7
13	No	1	2
16	No	1	1
17	No	1	3

The data from Table 4 show more participants in the low acquisition group comitting disuse errors in the 100% block of trials with three participants comitting multiple errors. While there were four participants within the no acquisition group that comitted a disuse error, they all comitted one error only and that error occurred relatively early in the trial block. Although not statistically supported, the trends depicted in Table 4 suggest that the low acquisition group was more prone to errors within the 100% block of trials.

Due to the near ceiling performance of all three groups during the 100% block trials, further analysis of misuse and disuse errors will focus only on the 70% block trials. Figure 11 shows the mean misuse errors committed by group during the 70% block, $F(2, 57) = 73.069$, $p<.0001$. Post hoc t-tests using Tukey's HSD are shown in Table 5. These analyses show that the three groups differed significantly with respect to the number of misuse errors committed during the 70% block trials. The no acquisition group committed the most errors, $M = 5.5$, $SD = 1.0$. The most misuse errors possible within this block of trials was 6, indicating that most members of the no acquisition group continued to rely on the automation in the face of errors. The low acquisition group committed the next highest level of misuse errors, $M = 2.45$, $SD = 1.985$. The high acquisition group committed the fewest number of misuse errors, $M = .4$, $SD = .681$.

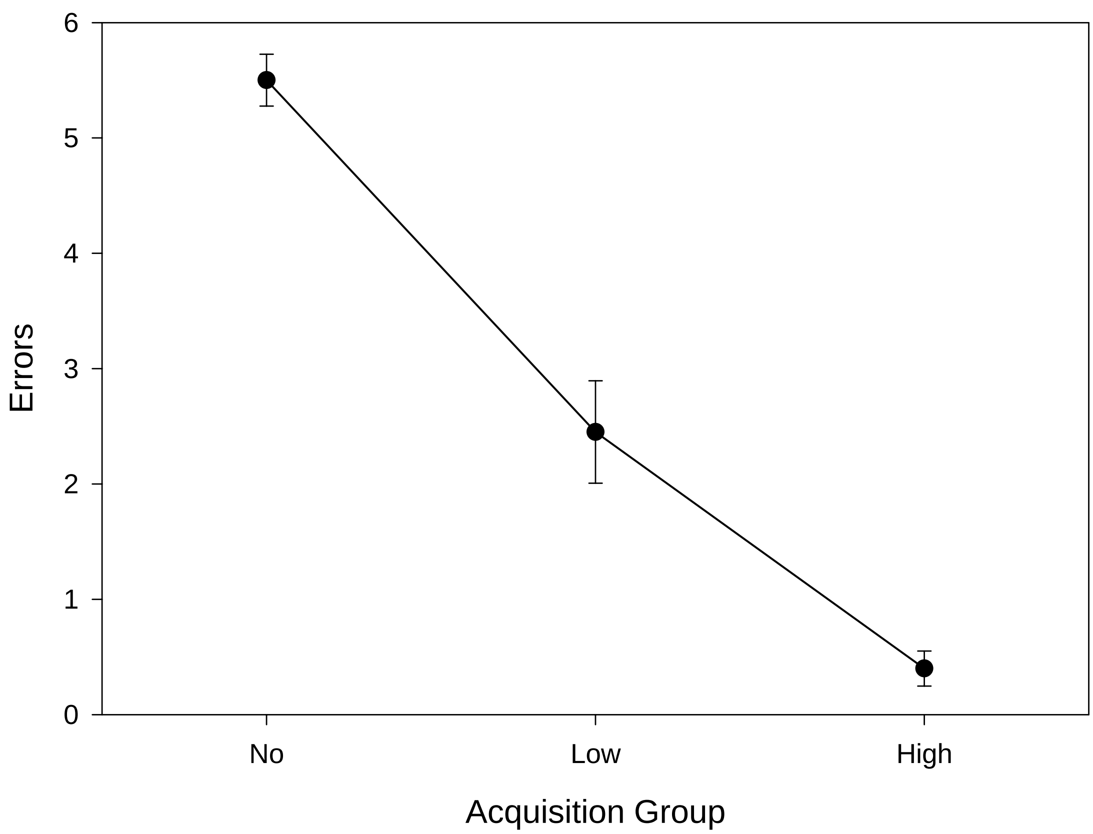

Figure 11. Misuse errors during 70% block trials.

Table 5. Post hoc independent groups t-tests for ANOVA of misuse errors

Trial Block	Groups	t	p
70% Automation Accuracy	No vs Low	6.13	<.0001*
	No vs High	-15.20	<.0001*
	Low vs High	-5.96	<.0001*

*Note.*Tukey's HSD, * Sig. at alpha = .05

Figure 12 shows the mean number of disuse errors committed by group during the 70% block. A univariate ANOVA of these means yielded significance, $F(2,57) = 6.615$, $p= .003$. To follow up this analysis, post hoc t-tests using Tukey's HSD were conducted and the results are displayed in Table 6. While the low acquisition group committed more disuse errors (M= 1.50, SD= 1.701), the no acquisition group (M=.45, SD= .945) and the high acquisition group (M=.25, SD= .550) committed the same number of disuse errors statistically.

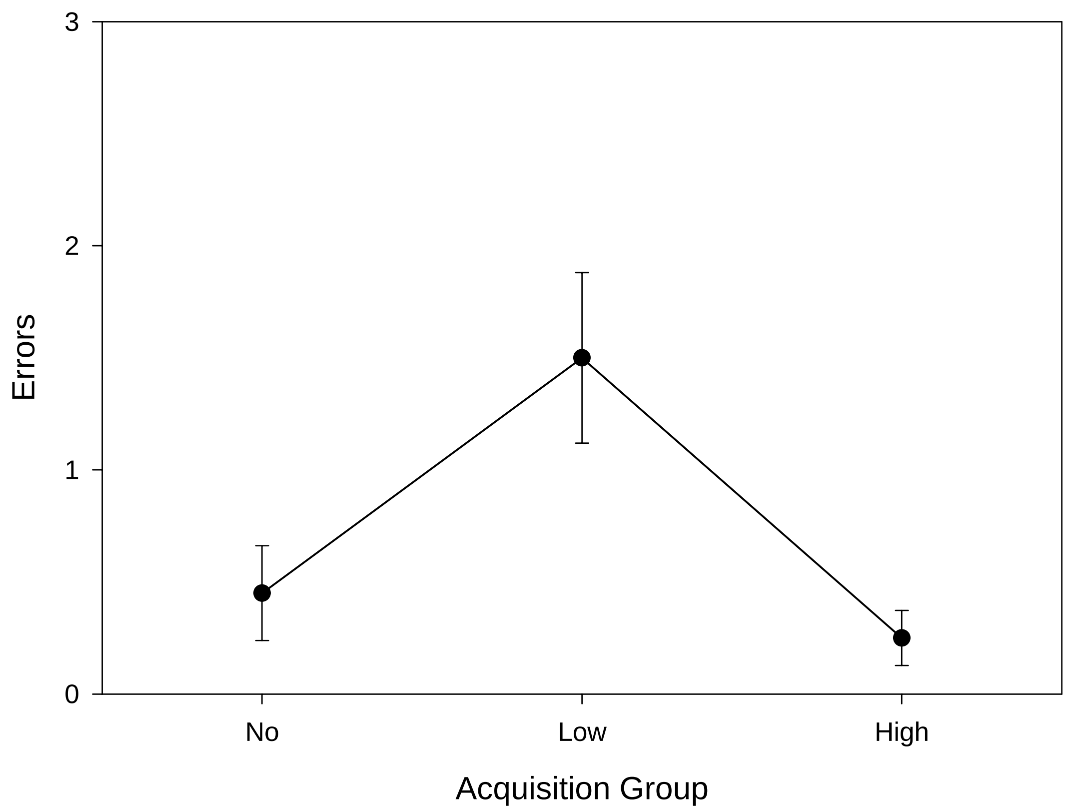

Figure 12. Disuse errors during 70% block trials.

Table 6. Post hoc independent groups t-tests for ANOVA of disuse errors

Trial Block	Groups	t	p
70% Automation Accuracy	No vs Low	-2.41	.017*
	No vs High	0.82	.851
	Low vs High	3.13	.004*

*Note.*Tukey's HSD, *Sig. p <.05

Estimation of Automation Reliability

Between each block of navigation trials, participants were asked the following question; “Please indicate, using a number, the reliability of the automated navigation system over the last block of trials. (Example: I think the navigation system was XX% reliable.)” Figure 13 shows the mean responses for this question for each group by trial block. Table 7 shows the descriptive statistics not represented by Figure 13.

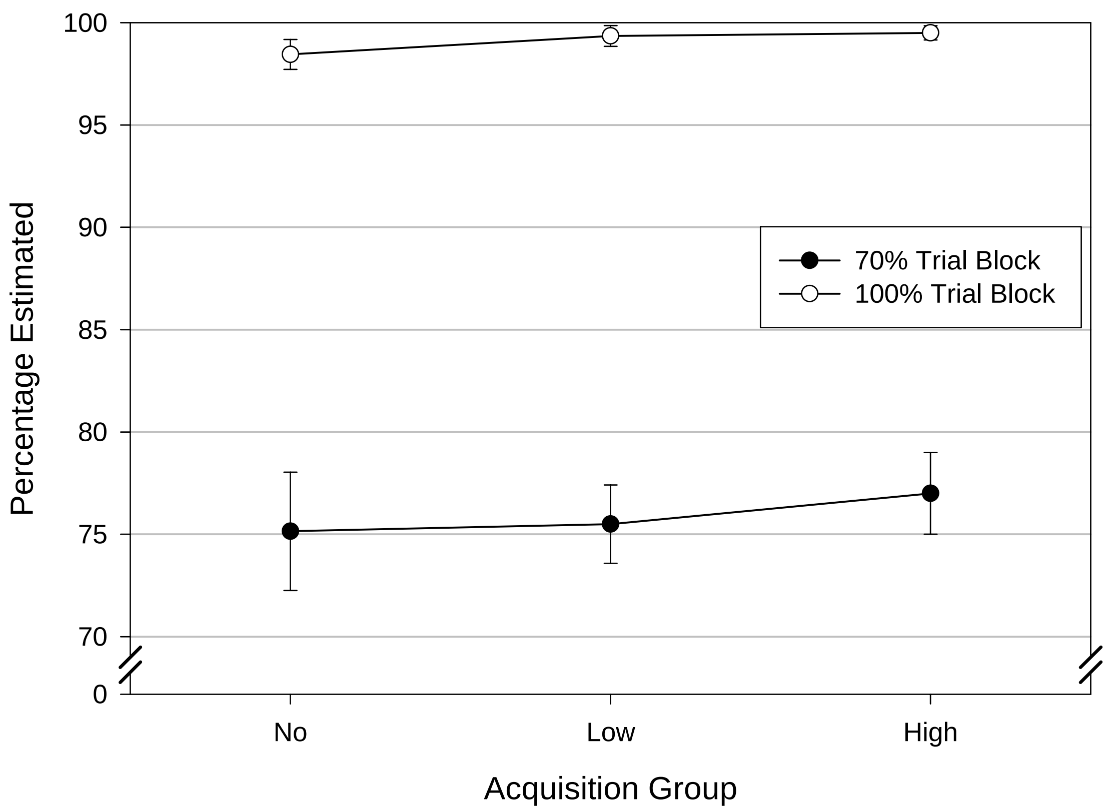

Figure 13. Group means of estimated automation reliability after each trial block.

Table 7. Descriptive statistics for automation reliability assessments post trial block.

Group		Range	Min	Max	M	SE	SD	Variance
No Acquisition	70% Assessment	48	50	98	75.150	2.892	12.934	167.292
	100% Assessment	10	90	100	98.450	0.731	3.268	10.682
Low Acquisition	70% Assessment	30	60	90	75.500	1.916	8.569	73.421
	100% Assessment	10	90	100	99.350	0.504	2.254	5.082
High Acquisition	70% Assessment	30	60	90	77.000	2.000	8.944	80.000
	100% Assessment	5	95	100	99.500	0.344	1.539	2.368

The results depicted in Figure 13 and Table 7 indicate no significant difference in the groups' ability to assess the reliability of the automation. During the 100% block trials, all three groups rated the reliability of the navigational aid near 100%. During the 70% block trials all three groups tended to overestimate the reliability of the aid. F-tests for the equality of variances between the three groups were non-significant. Thus, the means and the variances for the subjective measure of automation reliability were not significantly different with one exception. Table 8 shows the results of the F tests for equality of variances of assessments between the three acquisition groups for each block of trials. Only the variance of the high and no acquisition groups in the 100% block trial differ in their variability.

Table 8. F test values for equality of variances comparisons

		Low Acquisition	High Acquisition
70% Automation Accuracy	No Acquisition	2.2785	2.0911
	Low Acquisition		1.0896
100% Automation Accuracy	No Acquisition	2.102	4.512*
	Low Acquisition		2.146

Note. all comparisons had df of 19; *p<.05

Test of Learning

The test of learning was placed at the end of the navigation trials to measure any learning that may have resulted due to the experience of navigating through the city. Table 9 depicts the performance of the three acquisition groups on the test of learning compared with each groups' performance on the acquisition phase prior to the navigation trials if applicable. Paired sample t-tests were conducted to examine if there was any change in performance between the acquisition phase and the test of learning.

This analysis shows no significant change in performance, and presumably mental model, from the last acquisition phase and the test of learning for the high acquisition group. Conversely, the low acquisition group showed significant improvement from acquisition to test of learning. Since the no acquisition group started the study with the navigation trials, there are no descriptive statistics for the acquisition iteration and no paired sample t-test results for this group.

Table 9. Mean and Standard deviation statistics by group for last acquisition iteration and test of learning. Paired sample t-tests if applicable.

	Acquisition Iteration		Test of Learning		Paired T-test	
Acquisistion Group	M	SD	M	SD	T	p
No	NA	NA	15.74%	10.22%	NA	NA
Low	33.03%	11.80%	55.85%	27.17%	-4.673	<.0001*
High	96.26%	3.72%	96.98%	2.97%	-0.649	0.524

Note. * p<.025

The test of learning serves as the best method for measuring mental model quality in this study. Figure 14 depicts performance during the 70% block of trials plotted against performance on the test of learning. A regression analysis yielded R^2= .571. Therefore, if the test of learning is an adequate representation of the participants' mental model, then it would seem that the quality of the mental model predicts performance. In other words, the more accurate the mental model the better the performance during navigation trials in the 70% block.

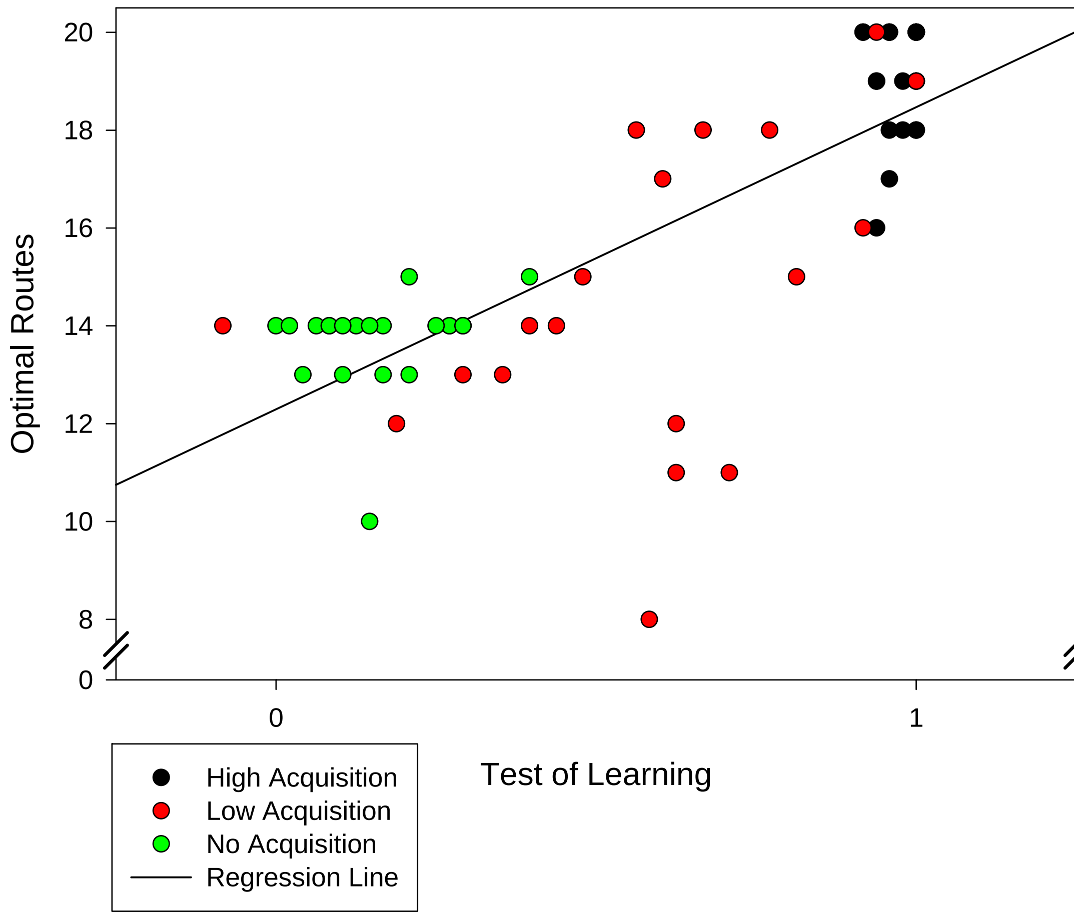

Figure 14. Test of learning performance vs. optimal routes selected during 70% block trials

Subjective Survey

The nine question subjective survey was originally designed to assess the participants' self-confidence in navigation ability, feeling of team interdependence, and trust in the automation. However, after running a principal components analysis on the survey data, only two dominant components were found. The principal components analysis was followed up with an exploratory factor analysis on the two dominant components. The loadings of this analysis suggested that the two common factors present in the data were representations of self confidence and trust. Table 10 displays the loadings of the common factors for each of the nine questions in the subjective survey. Questions 1, 3, and 7 loaded primarily on the common factor of self-confidence whereas questions 2, 4, 5, 6, 8, and 9 loaded primarily on the common factor of trust in the automation. Figure 15 shows the factor plot in rotated factor space depicts the dimensionality of the items within the survey.

*Table 10.*Rotated factor matrix for subjective survey

	Common Factors	
	Confidence	Trust
Survey Question	1	2
1	0.818	0.056
2	-0.068	0.688
3	0.825	0.113
4	0.100	0.727
5	0.066	0.781
6	0.060	0.529
7	0.810	0.001
8	-0.588	0.508
9	-0.065	0.432

*Note.*Principal axis factoring with Varimax rotation

Figure 15. Factor plot in rotated factor space.

Measures for trust in the automation and self-confidence in navigation ability were created using the questions associated with each common factor. A univariate ANOVA showed the three groups differed significantly on the measure of trust, $F(2, 57) = 8.554$, $p = .001$. The trust means for each group are illustrated in Figure 16. Post hoc multiple comparisons using Tukey's HSD were performed and are depicted in Table 11. The low acquisition group's rating of trust in the automation was significantly higher than the ratings of the other two groups. Prior to the study, participants completed a general trust in automation survey (Jian, Bisantz, & Drury 2000) intended to measure their preconceived notions of automation and reliability. The three groups did not differ in their ratings of trust in automation prior to the study; $F(2,57)= 1.231$, $p=.30$.

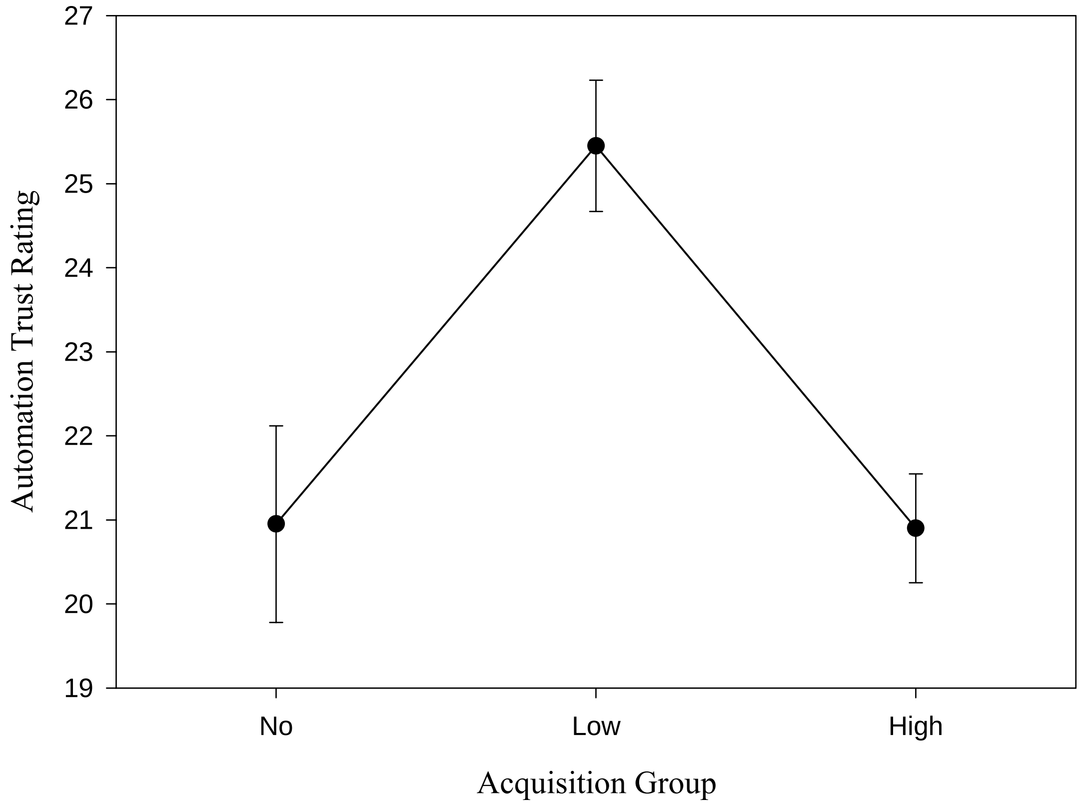

Figure 16. Mean trust scores per group.

Group means for the self-confidence measure were also tested using an ANOVA, $F(2,57) = 38.156$, $p < .0001$. Post hoc multiply comparisons using Tukey's HSD (Table 11) showed the high acquisition group as having significantly higher scores than the no and low acquisition groups. There was no significant difference between the confidence scores for the no and low acquisition groups. Figure 17 depicts the means of the three groups on the self-confidence measure.

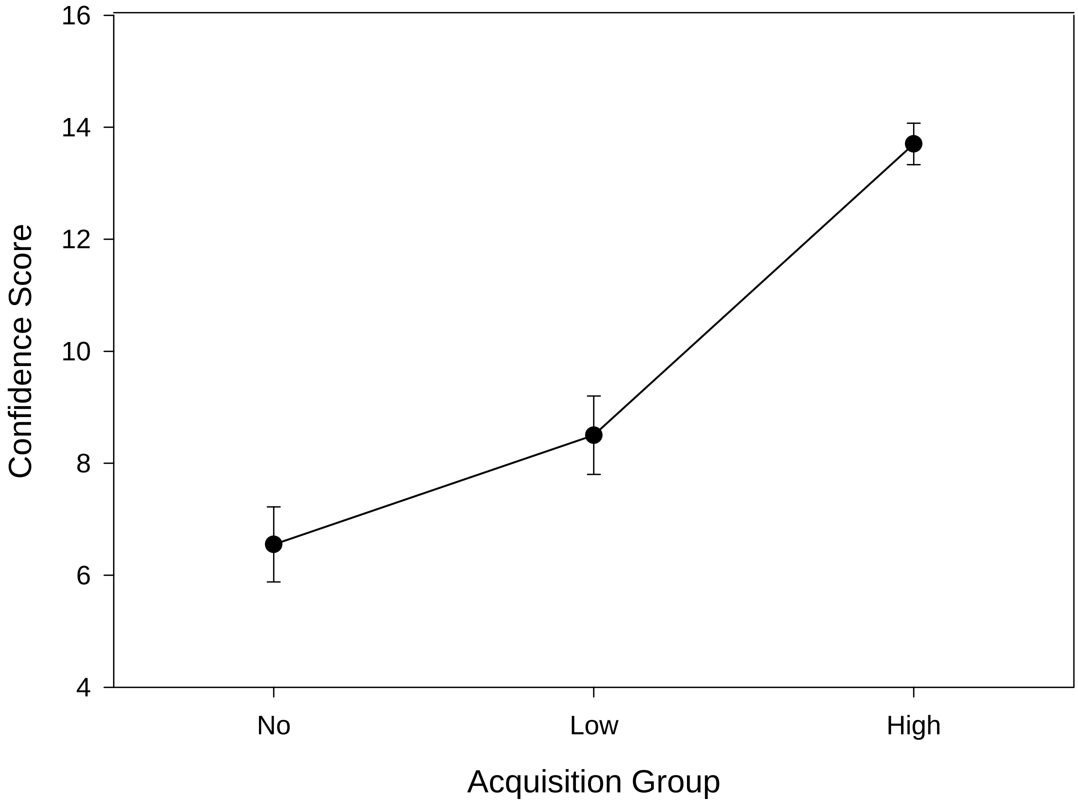

Figure 17. Mean confidence score by group.

Table 11. Multiple comparisons for Trust and Self Confidence.

		Low Acquisition	High Acquisition
Trust	No Acquisition	.002*	.999
	Low Acquisition		.002*
Self Confidence	No Acquisition	.063	<.0001*
	Low Acquisition		<.0001*

Note. *p<.05 Tukey's HSD

To evaluate the reliability of the common factors of trust and self confidence Cronbach's alpha was used. The statistic measures how well a set of items measure a one-dimensional latent construct and is often thought of as a coefficient of reliability. Table 12 displays Cronbach's alpha for all nine items within the subject survey as well as Cronbach's alpha for the common factors of trust and self confidence. The relatively high coefficients for trust and self confidence compared with the alpha for all items supports the claim that the items within each common factor are measuring the same construct; trust and self confidence.

Table 12. Reliability coefficients for common factors.

Items	Construct	Cronbach's Alpha
1 through 9	Survey	0.651
1, 3, & 7	Self Confidence	0.857
2, 4, 5, 6, 8, & 9	Trust	0.759

CHAPTER 4

DISCUSSION

This study was designed to answer two research questions; does quality of the user mental model predict performance when collaborating with automation and how does having a weak or vague mental model compare with having no mental model at all. These questions are important as they examine the effects of a variable (mental model) previously unstudied in association with automation.

Mental Models and Performance

The results of this study support the idea that mental model quality predicts performance while collaborating with an automated teammate. The high acquisition group performed significantly better than the other groups in navigating the city map and the no acquisition group predominately relied on the automation regardless of reliability. The low acquisition group, while performing at the same level as the no acquisition group in both trial blocks, showed significantly more variability in performance. In other words, some participants within the low acquisition group were able to acquire higher quality mental models while others were not (See Figure 14). The no acquisition group consistently complied with the recommendation of the automation and thus performed very well in the 100% block of trials but very poorly (matching the performance of the automation) in the 70% block of trials. Performance on the test of learning (see Figure 14) demonstrates the positive linear relationship between mental model quality and performance, regardless of group membership. Thus, users possessing high quality, accurate mental models of the task environment outperformed users possessing lower quality, inaccurate mental models.

Denying the no acquisition group's exposure to the task environment before navigation trials ensured participants within this group had no mental model to reference during navigation. The test of learning showed that this group acquired some knowledge during navigation; however, even with some learning the no acquisition group's mental models after navigation were worse than the low acquisition group's mental models prior to navigation. Yet, these two groups' performances, measured by optimal routes selected, were the same statistically. During the 100% block of trials, both groups performed near ceiling, but during the 70% block of trials they both performed near chance (relying solely on the automation in the 70% block of trials would result in 14 optimal routes selected). Whereas the two groups performed equally, the number and type of errors committed were significantly different. The low acquisition group committed significantly more disuse errors and significantly fewer misuse errors than the no acquisition group. Thus, in terms of performance, there was no difference between having some mental model and having no mental model, but the type of errors committed by each group was very different.

Whereas the no acquisition group's performance was certainly not good (chance level), it was predictable. It follows that a participant with no mental model of the task environment would rely on an aid even in the face of errors. The advantage of this predictability in the applied setting is if one knows the likely error rate of the automation one could mitigate the number of errors committed by an inexperienced (no mental model) user by adding supervision, adding systems, or by taking some other action to mitigate the risk of errors by the system. The low acquisition group's performance, equal to that of the no acquisition group, was unpredictable. Meaning the group committed both disuse and misuse errors. Disuse errors are potentially worse than misuse errors because they are the fault of the user only. The automation performs correctly, but the user rejects the recommendation and fails in spite of it. In the applied setting this could manifest itself in the mildly experienced user deviating from the correct recommendation of the automation and committing egregious system errors. This type of user possesses a mental model that may be incorrect and therefore conflict with the automation even when the aid is correct. The unpredictable nature of the low acquisition group supports the claim that having a weak or vague mental model may be worse than having no mental model at all; supporting the anecdotal saying "just enough information to be dangerous".

Self-Confidence and Trust

Numerous studies have proposed that users posses a bias towards automation that results in a perfect automation schema (Dzindolet, Peirce, Beck, Dawe, & Anderson, 2001; Dzindolet, Peirce, Beck, & Dawe, 2002; Madhavan, Weigmann, & Lacson, 2006; and Weigmann, 2002). This schema holds that in general, automated systems operate without error. Another theorized result of the perfect automation schema is errors identified as easy degrade trust rapidly (Madhaven et al.). In the current study, the ability to identify automation errors as easy or hard varied with group. For example the high acquisition group, having studied and been tested on the map numerous times, was better equipped to classify automation errors. The low acquisition group, having some prior map experience, was somewhat equipped to classify automation errors. It would follow then, that the errors produced by the automation were perceived as easy errors by the high acquisition group and this perception severely degraded the trust held for the automation. The low ratings of trust (compared to the low acquisition group) support this.

Because the no acquisition group experienced the city map through the navigation task alone, they were not equipped to classify automation errors, but they were able to identify when the automation made an error. The no acquisition group had little choice but to comply with the recommendation of the automated aid. Correspondingly, this group made more misuse errors and very few disuse errors. By following the recommendation of the automation in almost every instance this group was able to realize the true frequency of automation generated errors without having the mental model quality that the high acquisition group possessed. Thus, their group ratings of trust were significantly lower than those of the low acquisition group, but statistically equal to those of the high acquisition group.

Why were the low acquisition group's ratings of trust so high? Measured strictly on performance (optimal routes selected), the no and low acquisition groups were equal. The difference was in the type of errors committed. As previously stated, the no acquisition group generally complied with the recommendation of the automation and predominately committed misuse errors. The low acquisition group committed the same total number of errors, but committed significantly fewer misuse errors and significantly more disuse errors. Due to the nature of the task, a misuse error always resulted in the participant knowing that the automation was responsible for the error. However, a disuse error potentially created some doubt as to which team member was responsible for the error in navigation. In the process of making a disuse error, the participant viewed the recommendation of the automation (route depicted by the green line), decided it was inaccurate, and then selected a different route. Upon receiving the feedback "Sorry, you ended up in the wrong place." the participant knew their route was incorrect, but they potentially could not know if the automation's route was incorrect. Thus, disuse errors probably resulted in lowering the self-confidence of the participant while increasing the user's trust in the automation. Under the perfect automation schema, hard to identify errors result in higher levels of trust, and not surprisingly, the low acquisition group had the highest level of trust in the automation.

Self-confidence has been identified as an important factor in automation compliance behavior. Lee and Moray (1992, 1994) found that when user self-confidence exceeded the perceived reliability of the automation, then the user would select manual operation. Conversely, if the self-confidence of the user fell short of the perceived reliability of the automation, then the user would comply with the recommendation of the automation. Other studies, focusing on navigation systems and driver acceptance behavior found similar trends (Bonsall, 1992; Bonsall & Joint, 1991; Bonsall & Parry, 1990). Bonsall and colleagues found that drivers would ignore navigation systems once they entered into a familiar part of town, but consulted the aid when they entered unfamiliar sections of a city. Correspondingly, participants in the high acquisition group possessed high levels of self confidence and their performance on the test of learning shows they were familiar with the entire city. Thus, they most likely relied on their own abilities to navigate the city and discarded the recommendation of the automation. Additionally, because they knew the city well, automation errors would be classified as easy errors and their trust in the automation would be degraded. The low acquisition group possessed lower levels of self confidence and judging from their performance on the test of learning were unfamiliar with certain areas of the map. They possessed higher levels of trust and committed both misuse and disuse errors. Their misuse errors could be labeled as automation errors, but the disuse errors would create some ambiguity as to the true reliability of the aid. Thus the low acquisition group's level of trust was higher than that of the high acquisition group while the level of self confidence was lower.

Implications

The results of this study obviously imply that a collaborative system (team of human and automated system) will perform at a much higher level if the user's mental model is accurate or more complete. Getting users to a level where they feel confident in their own abilities, yet understand the abilities and limitations of the automation will take time and resources. This extra cost may make the added training prohibitive to some organizations, but the alternative is even less desirable. Putting users through crash course training sessions and hoping their mental models develop during on the job training could result in unpredictable performance by the collaborative system; a performance that includes errors by the automation (misuse) and errors by the user (disuse). Although organizations and businesses might not think of putting a user in a collaborative system without training, the results of this system in terms of pure performance would be comparable to that of a collaborative system containing a user with a weak mental model produced by substandard training.

Caveats

One limitation of this study is the limited scope of the team mental model studied. Cannon-Bowers, Salas, and Converse (1993) identified four types of team mental models and placed the environment (simulated city map in this study) as one of the aspects of the job/task mental model. To understand fully how the collaborative system team is affected by changes in mental model, other types and aspects of mental models need researching. For instance, another type of team mental model deals with the technology equipment. In a collaborative system, where one of the team members is the technology equipment, this type of mental model is crucial. The likely failures of the automation, the programming behind it, and sources of information used by the aid all combine to form this type of mental model.

Another limitation of the study was the fidelity of the task environment. While it was important to control the task environment for this study, the generalizability of the results to actual navigation with an in-car system might suffer from lack of fidelity. There are possibly immeasurable variables affecting the performance of a collaborative system made up of a driver and an in-car navigation system attempting to navigate the confines of a city. A logical follow on study would attempt to replicate these results in a higher fidelity environment (driving simulator) or actual driving environment to identify some of these other variables and their effects on performance.

The measurement of mental models is also critical for extending this research. The acquisition tasks and test of learning, similar to those used by Gilbert and Rogers (1999), provide an approximation of the knowledge possessed by the participants about the map. Whether participants used that knowledge as a mental simulation to describe, predict and explain the task environment as they navigated through it was not directly assessed. Yet, given only the blank city map, an origin, and a destination, the participants were forced to determine an optimal route using some knowledge structure and the mental model construct most likely approximates that structure.

A follow on study could incorporate a minor change to determine the extent of participant usage of the automation. In its current form, the navigation trial displays the recommendation of the automation for the duration of the trial. The green line is displayed and available for the participant to reference as desired. The ambiquity comes from not knowing if this information is used and to what extent it’s being used by the participants from each acquisition group. A simple change hiding the green line then displaying the route when the participant demands it (i.e., presses the space bar) would indicate the actual reliance behavior of the participant. The work of Bonsall (1990, 1991, &1992) and Lee and Moray (1992 & 1994), suggests that participants will reference the automation as a function of their self-confidence. For instance, members of a high acquisition group would likely reference the automation occasionally, but members of the low acquisition group would have higher levels of reliance and the members of the no acquisition group would likely reference the aid on every trial. This relativley simple change could potentially reveal more about the relationship between mental models and trends for automation usage. Figure 18 depicts a qualitative model that may describe the variables affecting automation usage. The model shows that mental model quality directly affects self confidence and the ability to classify and identify automation errors. As previously noted, when self confidence exceeds perceived reliability or trust users will most likely rely on their own abilities to complete a task. Under the perfect automation schema errors classified as easy tend to degrade trust rapidly. By incorporating the simple change described (automation recommendation on demand), a follow-on study should be able to test the accuracy of this qualitative model.

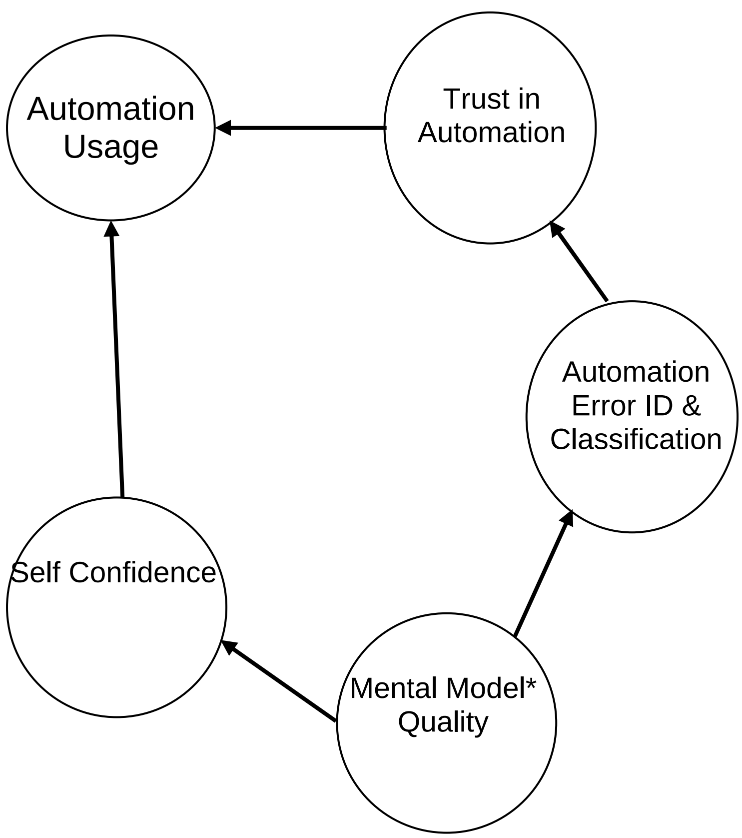

Figure 18. Qualitative model of mental model effects on automation usage. *Note.* * refers to mental model of the task environment in this specific study.

As Wilson and Rutherford (1989) put it, “There is, perhaps, a desire to apply it (mental models) to a whole range of issues even before we know if it has utility as an explanation of mind and behavior” (p. 618). While Wilson and Rutherford warned against the abuse of the mental model construct to explain a whole host of phenomena they admitted, “there is still much for the human factors profession to explore and utilize in the mental models concept” (p. 630). They go on to mention the need for the field to explore various tasks in a variety of systems and how mental models explain performance within this problem space. The current study explored a collaborative system in a navigation task and while the problem of accurately measuring and explaining the mental model construct may need further refinement, the study shows that mental models are an important part of automation research and deserve future investigation.

APPENDIX A

Checklist for Trust between People and Automation

Below is a list of statements for evaluating trust between people and automation. There are several scales for you to rate intensity of your feeling of trust, or your impression of the system while operating a machine. Please circle the number that best describes your feeling or your impression.

(Note: 1 = not at all; 7 = extremely)

1) Automated systems are deceptive

1 2 3 4 5 6 7

2) Automated systems behave in an underhanded manner

1 2 3 4 5 6 7

3) I am suspicious of automated system's intent, action, or outputs

1 2 3 4 5 6 7

4) I am wary of automated system

1 2 3 4 5 6 7

5) An automated system's actions will have a harmful or injurious outcome

1 2 3 4 5 6 7

6) I am confident in automated systems

1 2 3 4 5 6 7

7) Automated systems provide security

1 2 3 4 5 6 7

8) Automated systems have integrity

1 2 3 4 5 6 7

9) Automated systems are dependable

1 2 3 4 5 6 7

10) Automated systems are reliable

1 2 3 4 5 6 7

11) I can trust automated systems

1 2 3 4 5 6 7

12) I am familiar with automated system

1 2 3 4 5 6 7

APPENDIX B

LIKERT SCALE SURVEY

Please consider all of the navigation trials you completed while answering these questions.

1. I was confident in my ability to navigate through the city.

Strongly Disagree		Neither agree nor disagree		Strongly Agree
1	2	3	4	5

2. The automated aid was correct most of the time.

Strongly Disagree		Neither agree nor disagree		Strongly Agree
1	2	3	4	5

3. I felt unprepared and unqualified to conduct the navigational task.

Strongly Disagree		Neither agree nor disagree		Strongly Agree
1	2	3	4	5

4. The automated aid hurt my performance more than it helped.

Strongly Disagree		Neither agree nor disagree		Strongly Agree
1	2	3	4	5

5. I felt that my teammate and I worked well together.

Strongly Disagree		Neither agree nor disagree		Strongly Agree
1	2	3	4	5

6. The automated aid was not my teammate.

Strongly Disagree		Neither agree nor disagree		Strongly Agree
1	2	3	4	5

7. The navigational task was very difficult.

Strongly Disagree		Neither agree nor disagree		Strongly Agree
1	2	3	4	5

8. I'm glad I had a teammate to help me when I wasn't sure about directions.

Strongly Disagree		Neither agree nor disagree		Strongly Agree
1	2	3	4	5

9. The automated aid was wrong more that I thought it would be.

Strongly Disagree		Neither agree nor disagree		Strongly Agree
1	2	3	4	5

APPENDIX C

EXAMPLES OF TRIALS

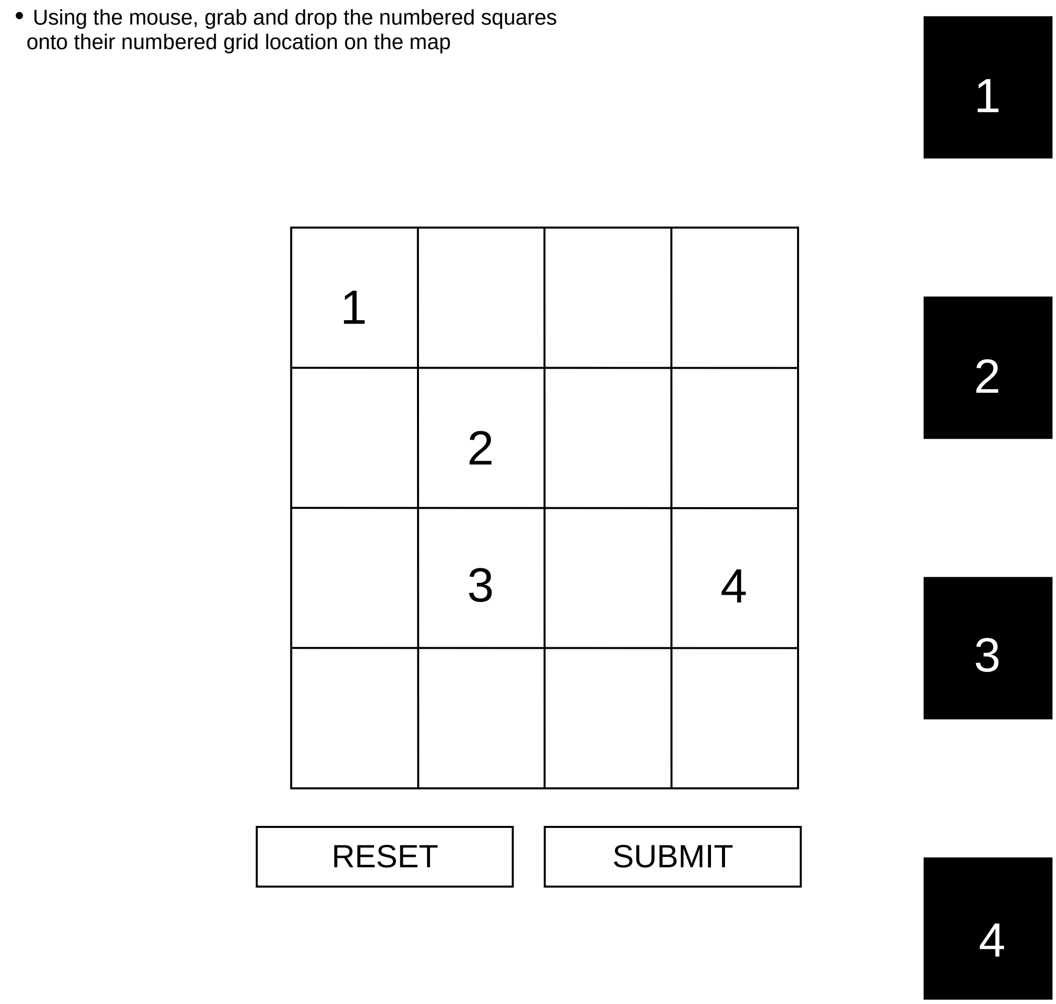

Figure 19. Example of drag and drop practice

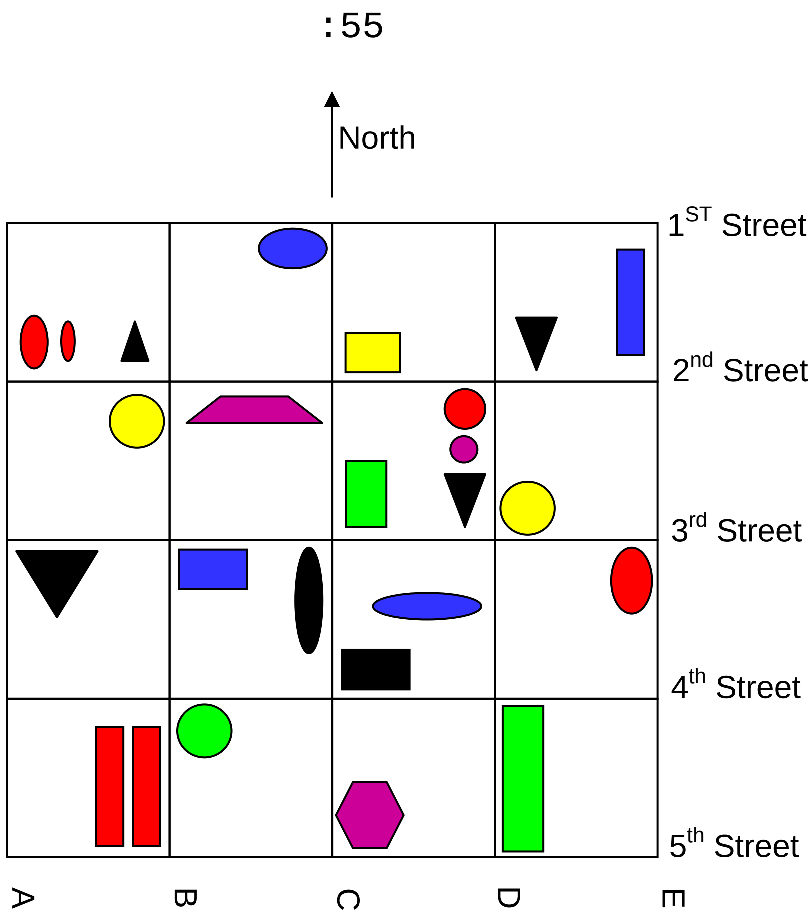

Figure 20. Example of map study.

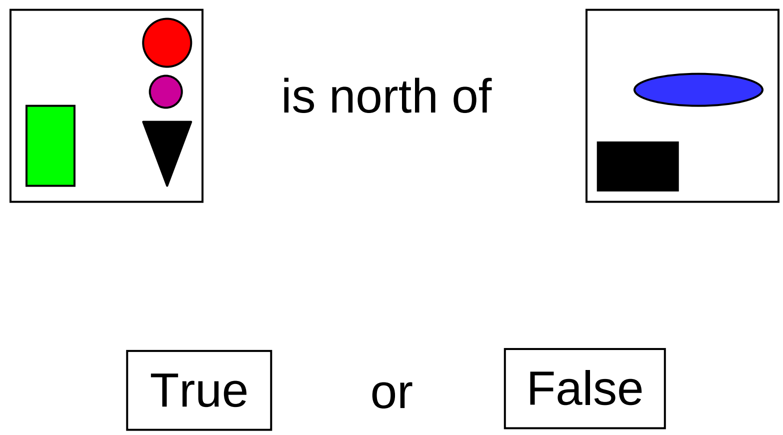

Figure 21. Example of relational question.

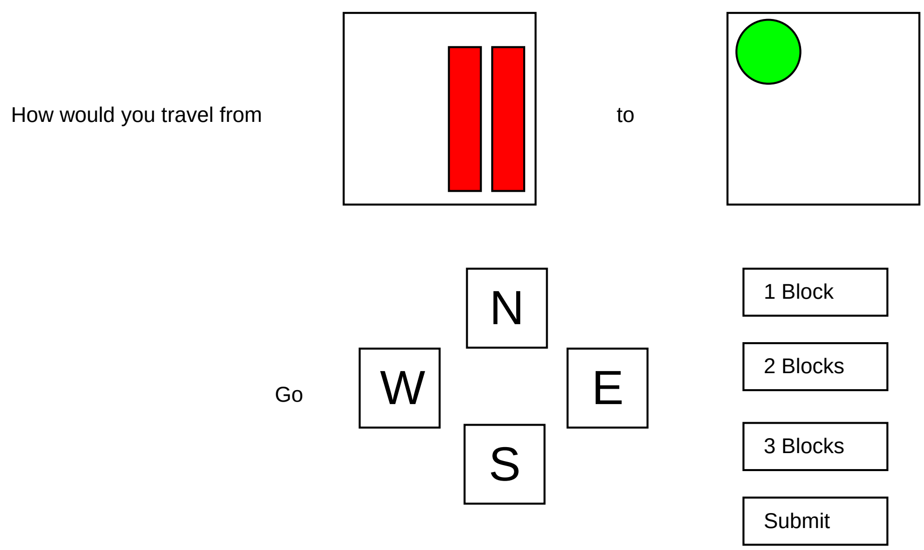

Figure 22. Example of directional question.

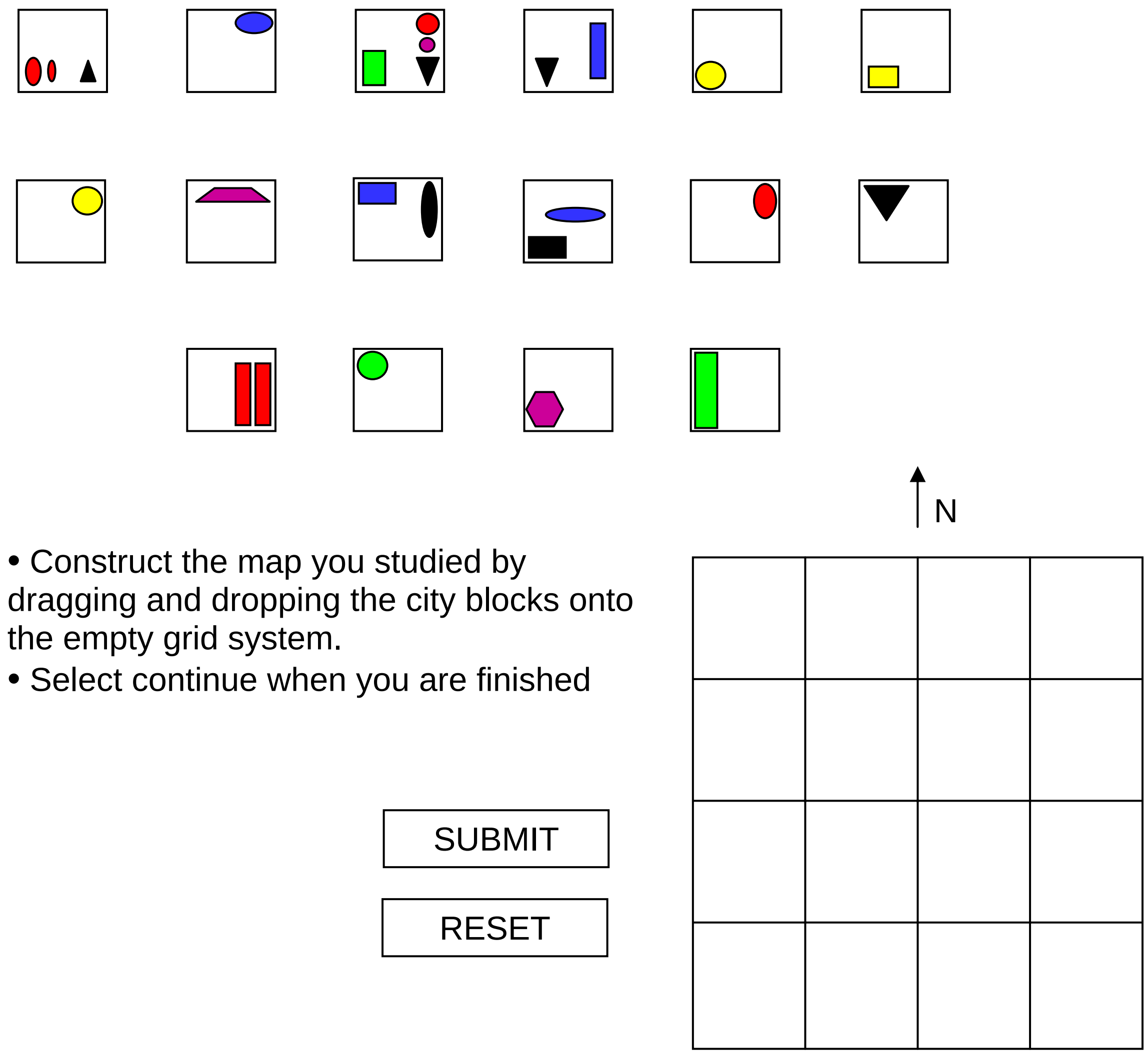

Figure 23. Example of map reconstruction task.

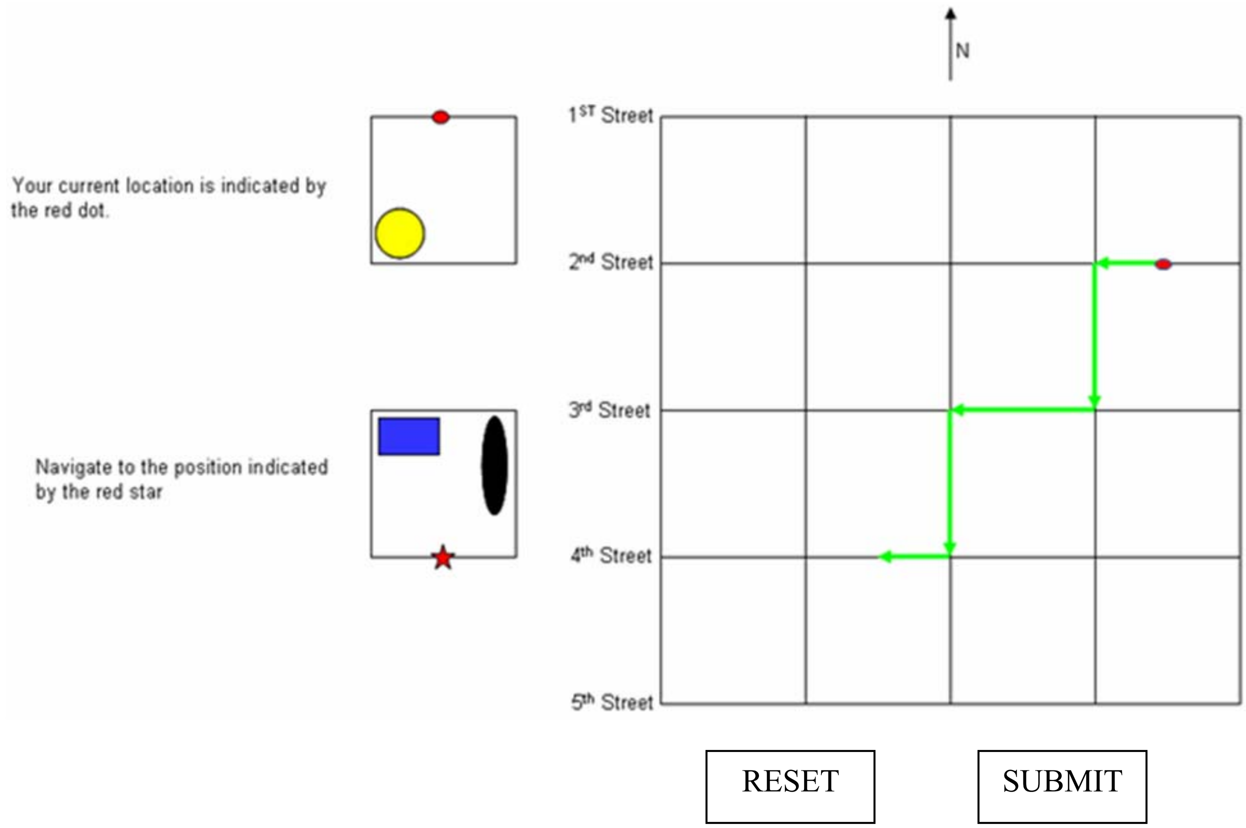

Figure 24. Example of navigation task.

APPENDIX D

DIRECTIONS FOR TASKS

Directions for Drag and Drop Practice (Mouse Practice)

● In each trial you will be required to move numbered boxes from the edges of the screen to their numbered locations on the grid system.

● To grab a box, move the cursor over it with the mouse and hold the left mouse button down.

● To drop a box, move the “grabbed” box over the desired square on the grid system and release the left mouse button.

● Once you have dragged and dropped all the numbered boxes onto the grid, select SUBMIT and continue onto the next trial.

Directions for Studying Map

● In the next trail you will be shown a 4 block X 4 block replication of a city.

● You will have one minute to study the city. Pretend you are looking at the city from above, with the top of the screen being north and the right side of the screen being east.

● Concentrate on the city blocks as you will be tested on your ability to remember the city by its blocks.

● When you are ready to continue select the CONTINUE button with the mouse and your one minute will begin.

Directions for Map Reconstruction

● In the following trial you will be required to reconstruct the city you just studied.

● The blocks of the city will be randomly arranged on the edges of the screen.

● Drag and Drop the city blocks onto the grid system in the same way that you practiced earlier.

● It is important to be as accurate as possible, but work quickly.

● When you are ready to proceed select the CONTINUE button.

Directions for Relational Questions

● In the following trials you will be required to answer a series of relational questions about the city you studied.

● You will see two city blocks with a relational statement.

● You must decide if the statement is true or false by selecting the appropriate button on the screen with your mouse. You will complete two practice trials followed by 16 test trials. Be as accurate as possible, but work quickly.

● When you are ready to proceed select the CONTINUE button.

Directions for Directional Questions

● In the following trials you will be required to answer a series of directional questions about the city you studied.

● You will see two city blocks and determine which cardinal direction you must travel for how many of blocks to get from one city block to the other. You will complete two practice triasl followed by 16 test trials. Be as accurate as possible, but work quickly.

● When you are ready to proceed select the CONTINUE button.

Directions for Navigation Task

● In the following trials you will navigate the city you have studied.

● You will see a blank city grid with labeled streets and avenues.

● You will be shown pictures of two blocks you studied. One will be your current location and one will be your desired destination. Using the mouse, you must draw the optimal route from origin to destination.

● To assist you in your journey is an automated navigational teammate. Your teammate will provide you with its best estimate of how to travel from one location to the other and is highly accurate, but not perfect.

● Here are some rules to follow during your journey: you can travel only one block on a particular street before you must turn to another street. Put another way, you must make a turn at every intersection.

● You and your teammate will conduct 2 practice trials with feedback and 20 test trials, all with feedback.

Things to remember

● There is only one optimal route per trial.

● You and your teammate will be evaluated collectively on your ability to navigate correctly and efficiently through the city.

● When you are ready to proceed select the CONTINUE button.

Directions for Automation Reliability Assessment

Please indicate, using a number, the reliability of the Automated Navigation System over the last block of trials. (Example; I think the Navigation System was XX% reliable.)

REFERENCES

Bonsall, P. W. (1992). The influence of route guidance advice on route choice in urban networks. *Transportation,* 19(1), 1-13.

Bonsall, P. W. & Joint, M (1991). Driver compliance with route guidance advice: the evidence and its implications. *Vehicle Navigation and Information Systems Conference,* 2, 47-59.

Bonsall, P. W. & Parry, T. (1990). Drivers' requirements for route guidance. *Proceedings of the 2nd International Conference on Road Traffic Control.* IEEE: London.

Cannon-Bowers, J. A., & Salas, E. (1998). Individual and team decision making under stress: Theoretical underpinnings. In J. A. Cannon-Bowers & E. Salas (Eds.), *Making decisions under stress* (pp. 17 – 38). Washington, DC: American Psychological Association.

Cannon-Bowers, J. A., Salas, E., & Converse, S. A. (1993). Shared mental models in expert team decision making. In N. J. Castellan Jr. (Ed.), *Current issues in individual and group decision making*. Hillsdale, NJ: Erlbaum.

Dixon, S. R., Wickens, D., & McCarley, J. S. (2006). How do automation false alarms and misses affect operator compliance and reliance? *Proceedings of the Human Factors and Ergonomics Society*, *50th Annual Meeting*, 25-29.

Dominguez, C. (1994). Can SA be defined? In M. Vidulich, C. Dominquez, E. Vogel, and G. Mcmillan (Eds.), *Situation Awareness: Papers and annotated bibliography* (pp. 5-15; Report AL/CF-TR-1994-0085). Wright-Patterson Air Force Base, OH: Air Force Systems Command.

Dzindolet, M. T., Peirce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human Factors, 44*(1), 79-94.

Dzindolet, M. T., Peirce, L. G., Beck, H. P., Dawe, L. A., & Anderson, B. W. (2001). Predicting misuse and disuse of combat identification systems. *Military Psychology, 13*(3), 147-164.

Dzindolet, M. T., Peterson, S.A., Pomranky, R. A., Pierce L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, 58(6) 697-718.

Ekstrom, R. B., French, J. W., & Harman, H. H. (1979). Cognitive factors: Their identification and replication. *Multivariate Behavioral Research Monographs, 79*(2), 3-84.

Endsley, M. R., (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors,* 37(1) 32-64.

Ezer, N. (2006). Toward an understanding of optimal performance within a human-automation collaborative system: Effects of error and verification costs, Unpublised master's thesis. Georgia Institute of Technology, Atlanta, Georgia.

Gilbert, K. D., & Rogers, W. A. (1999). Age-related differences in the acquisition, utilization, and extension of a spatial mental model. *Journal of Gerontology*, 54B(4) 246-255.

Goldin, S. E., & Thorndyke, P. W. (1982). Simulating navigation for spatial knowledge acquisition. *Human Factors,* 24(4), 457-471.

Ishihara, S. (1960). *Tests for Color-Blindness* (15^{th} ed.). Tokyo: Handaya Company.

Jastrzembski, T. S., Roring, R. W., & Charness, N. H., (2006). Videoconferencing technology as environmental support for older adults. *Proceedings of the Human Factors and Ergonomics Society*, *50^{th} Annual Meeting*, 175-179.

Jian, J., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, 4, 53-71.

Kantowitz, B. H., Hanowski, R. J., & Kantowitz, S. C. (1997). Driver acceptance of unreliable traffic information in familiar and unfamiliar settings. *Human Factors*, 39(2), 164-176.

Klein, G., Woods, D. D., Bradshaw, J. M., Hoffman, R. R., & Feltovich, P. J. (2004). Ten challenges for making automation a "team player" in joint human-agent activity. *Intelligent Systems, IEEE*, 19(6), 91-95.

Lee, J. D. (2005). Human factors and ergonomics in automation design. In G. Salvendy (Ed), *Handbook of human factors and ergonomics* (3rd ed., pp. 1570-1591). Hoboken, NJ: John Wiley and Sons.

Lee, J. D., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine studies. *Ergonomics,* 10, 1243-1270.

Lee, J. D., & Moray, N. (1994). Trust, self confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies*, 40, 153-184.

Langan-Fox, J., Anglim, J., & Wilson, R. (2004). Mental models, team mental models, and performance: Process, development, and future directions. *Human Factors and Ergonomics in Manufacturing*, 14(4), 331-352.

Madhavan, P., & Wiegmann, D. A. (2005). Effects of information source, pedigree, and reliability on operators' utilization of diagnostic advice. *Proceedings of the Human Factors and Ergonomics Society*, *49^{th} Annual Meeting*, 487-491.

Madhavan, P., Weigmann, D. A., & Lacson, F. C. (2006). Automation failures on tasks easily performed by operators undermine trust in automated aids. *Human Factors,* 48(2), 241-256.

Mathieu, J. E., Heffner, T. S., Goodwin, G. F., Cannon-Bowers, J. A., & Salas, E. (2005). Scaling the quality of teammates' mental models: Equifinality and normative comparisons. *Journal of Organizational Behavior,* 26(1), 37-56.

Mathieu, J. E., Heffner, T. S., Goodwin, G. F., Salas, E., & Cannon-Bowers, J. A. (2000). The influence of shared mental models on team process and performance. *Journal of Applied Psychology,* 85(2), 273-283.

McGuirl, J. M., & Sarter, N. B. (2006). Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. *Human Factors,* 48(4), 656-665.

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56(1), 81-103.

Nass, C., Fogg, B. J., & Moon, Y. (1996). Can computers be teammates? *International Journal of Human Computer Studies*, 45, 669-678.

Nass, C., Moon, Y., & Carney, P., (1999). Are people polite to computers? Responses to computer-based interviewing systems. *Journal of Applied Social Psychology,* 29(5), 1093-1110.

Nass, C., Moon, Y., & Green, N. (1997). Are machines gender neutral? Gender-stereotypic responses to computers with voices. *Journal of Applied Social Psychology*, 27(10), 864-876.

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. *Proceedings of the CHI'94 conference*, 72-78.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2), 230-253.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transaction on Systems, Man, and Cybernetics – Part A: Systems and Humans*, 30(3), 286-296.

Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places.* New York: Cambridge University Press.

Salas, E., Kosarzcki, M. P., Tannenbaum, S. I., & Carnegie, D. (2005). Aligning work teams and HR practices. In R. J. Burke & C. L. Cooper (Eds.), *Reinventing human resource management: challenges and new directions* (pp. 133-149). New York: Rutledge.

Salas, E., Prince, C., Baker, D. P., & Shrestha, L. (1995). Situation awareness in team performance: Implication for measurement and training. *Human Factors,* 37(1), 123-136.

Sanchez, J. (2005). Human-automation interaction: Factors that affect human behavior and system performance. Unpublished preliminary examination. Georgia Institute of Technology, Atlanta, Georgia.

Sanchez, J., Ezer, N., Rogers, W. A., & Fisk, A. D. (2006). Estimating reliability of automated aids: Effects of age and system reliability changes. *Cognitive Technology*, 11(1), 5-13.

Serfaty, D., Entin, E. E., & Johnston, J. H. (1998). Team coordination training. In J. A. Cannon-Bowers & E. Salas (Eds.), *Making decisions under stress* (pp. 224-227). Washington, DC: American Psychological Association

Sheridan, T. B. (2002). *Humans and automation: system design and research issues*. Santa Monica, CA: John Wiley and Sons.

Shipley, W. C. (1986). *Shipley Institute of Living Scale*. Los Angeles: Western Psychological Services.

Stout, R. J., Cannon-Bowers, J. A., Salas, E., & Milanovich, D. M., (1999). Planning, shared mental models, and coordinated performance: An empirical link is established. *Human Factors.* 41(1), 61-71.

Thorndyke, P. W. & Hayes-Roth, B. (1982). Differences in spatial knowledge acquired from maps and navigation. *Cognitive Psychology,* 13, 560-589.

Thorndyke, P. W. & Stasz, C. (1980). Individual differences in procedures for knowledge acquisition from maps. *Cognitive Psychology,* 12, 137-175.

Volpe, C. E., Cannon-Bowers, J. A., & Salas, E. (1996). The impact of cross-training on team functioning: An empirical investigation. *Human Factors.* 38(1), 87-100.

Wechsler, D. (1997). *Wechsler Adult Intelligence Scale III*. (3rd Ed.). San Antonio, TX: The Psychological Corporation.

Weigmann, D. A. (2002). Agreeing with automated diagnostic aids: A study of users' concurrence strategies. *Human Factors, 44*(1), 44-50.

Wickens, C., Dixon, S. R., & Johnson, N. (2006). Imperfect diagnostic automation: an experimental examination of priorities and threshold setting. *Proceedings of the Human Factors and Ergonomics Society*, *50th Annual Meeting*, 210 – 214.

Wilson, J. R. & Rutherford, A. (1989). Mental models: Theory and application in human factors. *Human Factors,* 31(6) 617-634.